Towards a plant model for enigmatic U-to-C RNA editing: the organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort Anthoceros agrestis

Gerke, Philipp ; Szövényi, Péter ; Neubauer, Anna ; Lenz, Henning ; Gutmann, Bernard ; McDowell, Rose ; Small, Ian ; Schallenberg-Rüdinger, Mareike ; Knoop, Volker

Abstract: Hornworts are crucial to understand the phylogeny of early land plants. The emergence of 'reverse' U-to-C RNA editing accompanying the widespread C-to-U RNA editing in plant chloroplasts and mitochondria may be a molecular synapomorphy of a hornwort–tracheophyte clade. C-to-U RNA editing is well understood after identification of many editing factors in models like Arabidopsis thaliana and Physcomitrella patens, but there is no plant model yet to investigate U-to-C RNA editing. The hornwort Anthoceros agrestis is now emerging as such a model system. We report on the assembly and analyses of the A. agrestis chloroplast and mitochondrial genomes, their transcriptomes and editomes, and a large nuclear gene family encoding pentatricopeptide repeat (PPR) proteins likely acting as RNA editing factors. Both organelles in A. agrestis feature high amounts of RNA editing, with altogether > 1100 sites of C-to-U and 1300 sites of U-to-C editing. The nuclear genome reveals > 1400 genes for PPR proteins with variable carboxyterminal DYW domains. We observe significant variants of the 'classic' DYW domain, in the meantime confirmed as the cytidine deaminase for C-to-U editing, and discuss the first attractive candidates for reverse editing factors given their excellent matches to U-to-C editing targets according to the PPR-RNA binding code.

DOI: https://doi.org/10.1111/nph.16297

New Phytologist

Research

Towards a plant model for enigmatic U-to-C RNA editing: the organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort *Anthoceros agrestis*

Philipp Gerke[1], Péter Szövényi[2], Anna Neubauer[2], Henning Lenz[3], Bernard Gutmann[4], Rose McDowell[5], Ian Small[5], Mareike Schallenberg-Rüdinger[1] and Volker Knoop[1]

[1]Institut für Zelluläre und Molekulare Botanik (IZMB), University of Bonn, Kirschallee 1, 53115 Bonn, Germany; [2]Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstr. 107, 8008 Zürich, Switzerland; [3]IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany; [4]EditForce Inc., West Zone #429, Kyushu University, 744 Motooka, Nishi-Ku, Fukuoka 819-0395, Japan; [5]ARC Centre of Excellence in Plant Energy Biology, University of Western Australia at Crawley, Perth, WA 6009, Australia

Author for correspondence:
Volker Knoop
Tel: +49 228 73 6466
Email: volker.knoop@uni-bonn.de

Received: *29 June 2019*
Accepted: *20 October 2019*

New Phytologist (2019)
doi: 10.1111/nph.16297

Key words: *Anthoceros agrestis*, chloroplast DNA, DYW domain, mitochondrial DNA, PPR proteins, PPR-RNA binding code, reverse U-to-C RNA editing, RNA editing factors.

Summary

- Hornworts are crucial to understand the phylogeny of early land plants. The emergence of 'reverse' U-to-C RNA editing accompanying the widespread C-to-U RNA editing in plant chloroplasts and mitochondria may be a molecular synapomorphy of a hornwort–tracheophyte clade. C-to-U RNA editing is well understood after identification of many editing factors in models like *Arabidopsis thaliana* and *Physcomitrella patens*, but there is no plant model yet to investigate U-to-C RNA editing. The hornwort *Anthoceros agrestis* is now emerging as such a model system.
- We report on the assembly and analyses of the *A. agrestis* chloroplast and mitochondrial genomes, their transcriptomes and editomes, and a large nuclear gene family encoding pentatricopeptide repeat (PPR) proteins likely acting as RNA editing factors.
- Both organelles in *A. agrestis* feature high amounts of RNA editing, with altogether > 1100 sites of C-to-U and 1300 sites of U-to-C editing. The nuclear genome reveals > 1400 genes for PPR proteins with variable carboxyterminal DYW domains.
- We observe significant variants of the 'classic' DYW domain, in the meantime confirmed as the cytidine deaminase for C-to-U editing, and discuss the first attractive candidates for reverse editing factors given their excellent matches to U-to-C editing targets according to the PPR-RNA binding code.

Introduction

The phylogenetic placement of hornworts (Anthocerotophyta) among land plants (Embryophyta) is still contentious (e.g. Cox, 2018). A consensus seemed to have been established that hornworts are sister to vascular plants (tracheophytes), suggested by the gain of a shared mitochondrial intron absent in the other two bryophyte clades, the liverworts and the mosses (Groth-Malonek *et al.*, 2005). A hornwort–tracheophyte (HT) clade was subsequently well supported by concatenated, organellar 'phylogenomic' sequence data sets (Qiu *et al.*, 2006). Recent phylogenetic analyses using nuclear transcriptome data sets, however, suggest alternative scenarios for the phylogeny of early embryophytes (Puttick *et al.*, 2018).

One intriguing character that would provide a further molecular synapomorphy of the HT clade is 'reverse' U-to-C RNA editing in plant mitochondria and chloroplasts. Whereas C-to-U RNA editing is present in all major land plant clades, including the liverworts and the mosses, no evidence has ever been found for U-to-C editing in these two clades (Malek *et al.*, 1996; Freyer *et al.*, 1997; Steinhauser *et al.*, 1999; Rüdinger *et al.*, 2012). However, there is no doubt that reverse U-to-C RNA editing is abundantly present in hornworts (Yoshinaga *et al.*, 1996; Steinhauser *et al.*, 1999; Kugita *et al.*, 2003b), in ferns (Vangerow *et al.*, 1999; Guo *et al.*, 2015; Knie *et al.*, 2016), and, among lycophytes, at least in the order Isoetales (Grewe *et al.*, 2011).

The mechanism of C-to-U-type RNA editing is reasonably well understood, mainly owing to the characterization of many RNA editing factors in model systems such as the flowering plants *Arabidopsis thaliana* and *Oryza sativa* and in the moss model system *Physcomitrella patens* (Barkan & Small, 2014; Ichinose *et al.*, 2014; Schallenberg-Rüdinger & Knoop, 2016). By now, *c.* 80 site-specific RNA editing factors have been characterized, recently summarized in the database EdiFacts, an addition to the PREPACT service for the analysis of plant-type RNA editing (Lenz *et al.*, 2018). These site-specific RNA editing factors are unique RNA-binding pentatricopeptide repeat (PPR) proteins featuring additional carboxyterminal domains called E1, E2, and DYW (Cheng *et al.*, 2016). Their upstream arrays of PPRs in editing factors are of the 'PLS-type' featuring classical 35 amino acid P-type repeats

along with shorter (S-type) and longer (L-type) variants (Lurin *et al.*, 2004; Cheng *et al.*, 2016). PPR arrays are fundamental for specific binding to transcripts in a one-PPR-per-ribonucleotide manner, and the essentials of a PPR-RNA recognition code with the fifth (5) and the last (L) amino acid of P-type and S-type PPRs recognizing individual nucleotides in the RNA target have been identified (Barkan *et al.*, 2012; Takenaka *et al.*, 2013; Yagi *et al.*, 2013; Kobayashi *et al.*, 2019; Yan *et al.*, 2019).

Given its evident similarity to known cytidine deaminases, including important zinc ion (Zn^{2+})-binding motifs, the terminal DYW domain is the prime candidate to carry the enzymatic activity to convert cytidines into uridines (Salone *et al.*, 2007; Iyer *et al.*, 2011; Boussardon *et al.*, 2014; Hayes *et al.*, 2015; Wagoner *et al.*, 2015; Ichinose & Sugita, 2018; Oldenkott *et al.*, 2019). The upstream PPR stretch for RNA recognition linked *in cis* to a downstream E1, E2, and the DYW domain is evident for all editing factors in the model moss *P. patens*. Thanks to the simplicity of this plant model, all organelle editing sites in the moss have been assigned to their corresponding DYW-type editing factors (Ichinose *et al.*, 2013; Schallenberg-Rüdinger *et al.*, 2013; Sugita *et al.*, 2013; Ichinose *et al.*, 2014; Schallenberg-Rüdinger & Knoop, 2016). However, the setup of organelle RNA editing is evidently more complex in flowering plants, where truncated proteins require interactions with DYW domains supplied *in trans*, frequently mediated by extra helper proteins (e.g. NUWA and multiple organellar RNA editing factor (MORF)/RNA-editing factor interacting protein (RIP) proteins) in much more complex editosomes (Takenaka *et al.*, 2012; Bentolila *et al.*, 2012; Boussardon *et al.*, 2012; Sun *et al.*, 2013, 2015, 2016; Zehrmann *et al.*, 2015; Diaz *et al.*, 2017; Bayer-Császár *et al.*, 2017; Andrés-Colás *et al.*, 2017; Guillaumot *et al.*, 2017; Sandoval *et al.*, 2019).

In contrast to C-to-U editing, we have no idea yet about the mechanisms of reverse U-to-C RNA editing. The main reason is that the editomes of the aforementioned model systems and those of other flowering plants seem to be entirely devoid of U-to-C RNA editing (Edera *et al.*, 2018; Lenz *et al.*, 2018). We could not corroborate occasional reports of reverse RNA editing in angiosperms (P. Gerke, V. Knoop, unpublished findings) and consider it likely that U-to-C RNA editing is phylogenetically restricted to hornworts, lycophytes, and ferns. Hence, the investigation of reverse RNA editing calls for a new model organism from one of the latter three plant clades. For this purpose, we consider the hornworts to be the most attractive candidates, assuming that, independent of their exact phylogenetic position among the bryophytes, they are phylogenetically closest to the evolutionary origins of U-to-C RNA editing.

Towards that goal, we report here on the assembly of the organelle genomes of *Anthoceros agrestis*, on the accompanying transcriptome and editome studies, and on the first analyses of the vastly extended and surprisingly diversified nuclear gene family of 'DYW-type' PPR proteins in that hornwort. We speculate that U-to-C RNA editing has originated from the more ancient and widespread C-to-U editing, using the same mechanisms for RNA target recognition linked to a biochemical enzyme variant, possibly converting a deaminase into a transaminase. Given the likely earlier evolutionary origin of plant C-to-U RNA editing among land plants it is suggestive that PPR proteins remain at the core of target recognition also for sites of U-to-C editing. We present the first preliminary candidates for potential U-to-C RNA editing factors to be investigated in future functional studies in *A. agrestis* as an emerging new model system in plant molecular biology.

Materials and Methods

DNA/RNA extraction and sequencing

Both RNA and DNA were extracted from the *A. agrestis* BONN strain described previously (Szövényi *et al.*, 2015). DNA samples (100 ng) were used to prepare paired-end DNA-sequencing libraries using the Nextera XT library preparation kit (Illumina Inc., San Diego, CA, USA) and each sequenced on one-third of a MiSeq flow cell (250 bp) at the Functional Genomic Center Zurich (FGCZ) as described. Raw reads were assembled using the A5-MISEQ pipeline (Coil *et al.*, 2015), specially designed for paired-end MiSeq reads and small genomes. Raw DNA reads were deposited in the European Nucleotide Archive and are available under study accession no. PRJEB8683. To identify sites undergoing RNA editing in the organellar transcripts, we extracted total RNA from 4-wk-old gametophyte tissues as described (Szövényi *et al.*, 2015). RNA was processed using the RiboMinus Plant Kit for RNA-Seq (Thermo Fisher Scientific) to deplete ribosomal RNAs (rRNAs) and used to prepare a stranded RNA-sequencing (RNA-seq) library (TruSeq mRNA library kit) that was paired-end sequenced (150 bp) on 1/4th lane of an Illumina HiSeq4000 machine at FGCZ. Raw RNA-seq reads were deposited in the European Nucleotide Archive and are available under study (run) accession no. PRJEB33107 (ERR3383408).

Organelle genome assembly

Assembly of next-generation sequencing (NGS) raw sequence data (ERR771108) was carried out using MEGAHIT software (Li *et al.*, 2016) with stepwise increase of *k*-mer values up to 141 bp to assemble sequence contigs. Mitochondrial and chloroplast contigs were initially identified with BLASTN searches using available organelle genomes as queries. Mitochondrial contigs were characterized by MEGAHIT 'multi' values reflecting coverage between 105 and 337, whereas those of chloroplast origin reached higher values of up to 1947 for the inverted repeat (IR) regions. Gaps between contigs were due to difficult microrepeat or homopolymer sequences in intergenic regions, which were filled by targeted PCRs in both organelles. The chloroplast and mitochondrial genomes of *A. agrestis* were submitted under NCBI/GenBank accession nos. MK087646 and MK087647, respectively.

Determination of RNA editing events

DNA reads (ERR771108) and newly generated RNA reads were trimmed (settings PE phred33 ILLUMINACLIP:TruSeq3PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) with TRIMMOMATIC v.0.35 (Bolger *et al.*,

2014) and mapped against the organelle genomes using GSNAP (Wu *et al.*, 2016). JACUSA (Piechotta *et al.*, 2017) was used to determine RNA–DNA differences among the two mapping files generated. To identify RNA editing sites we set thresholds of coverage by at least 30 reads and an editing efficiency of at least 5%. Care was given to problematic cases like RNA editing close to exon–intron borders, mapping to pseudogene fragments, or mismapping to rRNA sequences from the other organelle. RNA editing was independently determined for selected cases by targeted reverse transcription (RT)-PCR as discussed later.

Identification of candidate RNA editing factors

An updated version of the PPR finder tool (http://ppr.130.95.176.97.xip.io/fasta/) based on the recent reassessment of PPR-types, E1, E2, and DYW domains (Cheng *et al.*, 2016), was used to identify proteins encoding those domains in the *A. agrestis* genome assembly. A total of 3089 PPR proteins were identified, of which 1464 were selected as being of an 'E+' type, revealing at least the beginning of a DYW domain with the characteristic PG box or variants thereof at its amino terminus. Amino acids at positions 5 and L (last) were extracted from the PPR repeats for evaluation of candidate targets using the core rules of the PPR-RNA recognition code (Barkan *et al.*, 2012).

Phylogenetic tree construction

PPR proteins were aligned with MAFFT (Kuraku *et al.*, 2013) followed by manual adjustment. An alignment region comprising 191 positions including the three C-terminal PPRs P2, L2, and S2, the E1 and E2 domains (Cheng *et al.*, 2016), and extending into the first 20 amino acids of the DYW domain was selected for phylogenetic analysis given the variable downstream truncations of many *Anthoceros* PLS-type proteins and to avoid noninformative similarities arising from homoplasies within the further upstream PPRs. The set of 1464 proteins initially identified was reduced to 1428 for phylogenetic reconstruction since 36 proteins (including three 'pure' DYW proteins) showed degenerations in the C-terminal domains. Maximum likelihood phylogenetic tree construction was done with IQ-TREE v.1.6.5 (Trifinopoulos *et al.*, 2016) using the JTT+F+G4 model identified as best-fitting substitution model with the implemented MODELFINDER (Kalyaanamoorthy *et al.*, 2017). Node reliability was determined from 1000 bootstrap replicates with ultrafast bootstrap approximation UFBoot (Hoang *et al.*, 2018).

PPR target prediction

PPR positions 5 and L were extracted for P and S-type PPRs and translated into weight matrices as input for the TARGETSCAN module recently implemented in PREPACT (Lenz *et al.*, 2018). Arbitrary numerical assignments for matches according to the PPR-RNA recognition code were essentially as outlined previously, but now extended with weights for purine or pyrimidine ambiguities should only position 5 but not position L match according to the code rules, hence resulting in the weight matrix shown in Table 1.

Table 1 Weights assigned to individual PPRs used as the input for the TARGETSCAN feature of PREPACT (Lenz *et al.*, 2018) to scan for candidate RNA targets of individual PPR proteins.

PPR-type	Pos. 5	Pos. L	Nucleotide identity weights (%)				Position weight (%)
			A	C	G	U	
P or S	T OR S	N	90	0	10	0	200
P or S	T OR S	D	10	0	90	0	200
P or S	T OR S	NOT (N OR D)	50	0	50	0	200
P or S	N	N OR S	0	60	0	40	100
P or S	N	D	0	30	0	70	100
P or S	N	NOT (N OR D OR S)	0	50	0	50	100
L	ANY	ANY	25	25	25	25	0

Searches for targets were performed using the TARGETSCAN option 'around known editing sites' for the newly determined organelle editomes. Initial scores (ISC) for a match between a PPR protein and a candidate target are the sum of percentages for the individual positions. The ISC values were divided by the respective matrix length (ml) to compensate for the length differences of PPR arrays. For the 'reverse' assignments of PPR proteins to a given editing site, the rank (Rk) among the top matches for a given protein was additionally considered to result in an ultimate 'score-of-fit' SOF = ISC/(ml × Rk). To test for statistical significance in the assignments of different DYW-types to C-to-U vs U-to-C editing sites, a one-proportion *Z*-test vs equal proportions (0.5) was conducted.

Results

The assembly of the complete chloroplast and mitochondrial genomes of *A. agrestis* from NGS data were straightforward given their stoichiometric dominance in the total DNA preparations. On average, the respective contigs representing single-copy sequences in the three different plant genomes had coverages of above 1000 for chloroplast sequences, of *c.* 100–170 for mitochondrial sequences and *c.* 5–20 for nuclear sequences.

The *Anthoceros agrestis* plastome

The chloroplast genome of *A. agrestis* is conserved as typical for land plants, featuring the canonical arrangement of a large single-copy (LSC) region and a small single-copy (SSC) region separated by a pair of inverted repeats (Fig. 1). The total chloroplast DNA (cpDNA) size is 160 760 bp, consisting of an LSC of 107 329 bp and an SSC of 22 167 bp separated by a pair of IRs of 15 632 bp each. Likewise, the *A. agrestis* cpDNA features an expected gene complement. Noteworthy characteristics, however, are a continuous, large *ycf1* reading frame, which is disrupted in the cpDNAs of *Anthoceros angustus*, formerly referred to as *Anthoceros formosae* (Kugita *et al.*, 2003a) and *Nothoceros aenigmaticus* (Villarreal *et al.*, 2013). A continuous *ycf1* is, however, also present in the

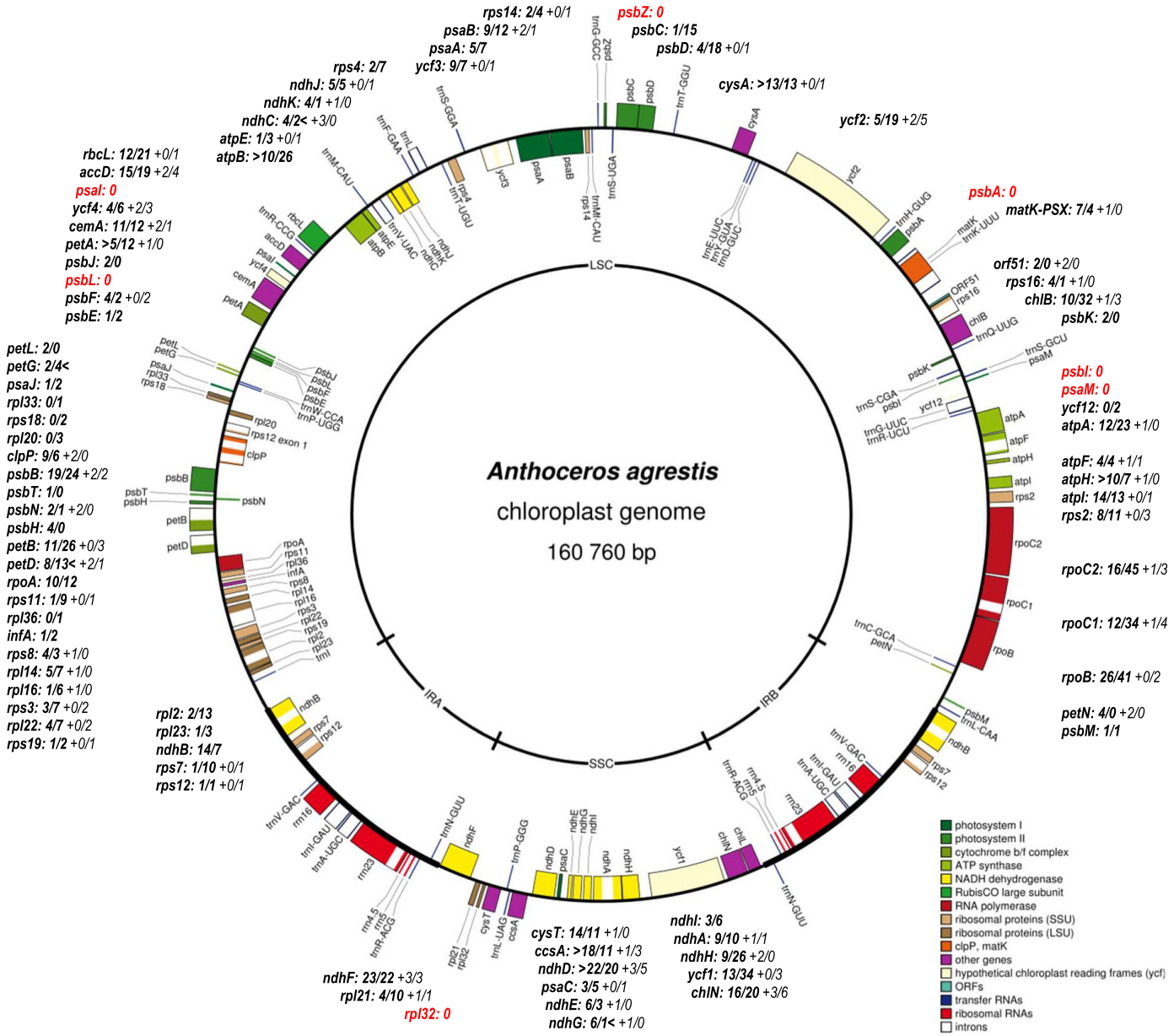


Fig. 1 The *Anthoceros agrestis* chloroplast genome. The chloroplast DNA (cpDNA) map was drawn with OGDRAW (Lohse *et al.*, 2013). The *A. agrestis* cpDNA (deposited in the database under accession no. MK087646) has a typical plant circular plastome structure consisting of a large and a small single-copy region separated by a pair of inverted repeats and the expected gene complement. Gene categories are indicated in the legend. Notable features are the presence of a *trnS-CGA* gene between *psbK* and *psbI*, a continuous long *ycf1* reading frame, and the presence of a group I intron in the rrn23 gene for the large ribosomal rRNA gene (*rrn23*). The numbers given for all protein-coding sequences indicate nonsilent C-to-U/U-to-C edits (bold) and additional silent edits after the plus sign. Symbols '>' and '<' indicate creation of start and stop codons by RNA editing, respectively. Genes indicated in red lack messenger RNA editing.

very recently determined cpDNA of *Leiosporoceros dussii* (Villarreal Aguilar *et al.*, 2018). A group I intron (rrn23i2620g1) in the *rrn23* gene for the large chloroplast rRNA is exclusively present in the genus *Anthoceros*. Both observations are noteworthy, given a more ancestral state of evolution and an extended intron complement, which we also observe for the mitochondrial DNA (mtDNA) in *A. agrestis* (see below). The IRs are extended to include the 3′-ends of *rps12*, *rps7* and *ndhB*, which are part of the LSC in other land plants, including *Leiosporoceros* and *Nothoceros*. As in the other hornwort plastomes, *A. agrestis* also lacks the *rps15* gene ancestrally located between *ycf1* and *ndhH* in the SSC. We detected a *trnS-CGA* gene between *psbK* and *psbI*, which is also present in *Leiosporoceros*. Upon closer inspection, we found that this peculiar *trnS* gene is also conserved in the liverwort *Pellia endiviifolia*, in the moss *Takakia lepidozioides*, and in the other hornwort plastomes but had previously been missed in the respective annotations.

The *Anthoceros agrestis* chloroplast editome

Most events of RNA editing in plant organelles serve to reconstitute conserved amino acid identities in protein coding

sequences. We predicted 1371 candidate sites of RNA editing using the 12 nonangiosperm chloroplast references available with the latest update of PREPACT and the default 'commons' threshold level of 70% (Lenz *et al.*, 2018). Analysing the transcriptome data, we ultimately identified 1549 sites of chloroplast RNA editing (636 C-to-U and 913 U-to-C edits) in the *A. agrestis* chloroplast (Supporting Information Table S1). We use the previously proposed nomenclature to designate RNA editing sites (Rüdinger *et al.*, 2009; Lenz *et al.*, 2010) indicating the affected gene, followed by 'eU' or 'eC' to indicate creation of uridine or cytidine, respectively, followed by its position and finally, for the dominating type of edits in coding regions, the effected codon sense change. Hence, edit 'atpAeU2TM' would create a methionine (AUG) start codon from a genomic (ACG) threonine codon in the *atpA* mRNA. The chloroplast editome includes 67 editing sites in 5′ and 3′ untranslated regions (UTRs), 27 editing sites in introns, and 124 silent edits in coding regions that could not be predicted. No case of silent editing in either direction of pyrimidine exchange was observed in AGY serine, CGY arginine, GGY glycine, or UGY cysteine codons, fully matching previous observations of only very rare RNA editing immediately downstream of a guanidine (Lenz *et al.*, 2018). Many silent sites and those outside of coding regions are edited to much lower degrees than those in codon-changing positions (Table S1).

The confirmed sites of nonsilent RNA editing fit the predictions very well. Altogether, 275 in-frame stop codons are removed from reading frames by U-to-C editing (Table S1) and six translation start codons (in *atpB*, *atpH*, *ccsA*, *cysA*, *ndhD*, and petA) and four stop codons (in *ndhC*, *ndhG*, *petD*, and *petG*) are created by C-to-U editing (Fig. 1). Only five very short reading frames and *psbA* are not affected by RNA editing (Fig. 1). The *atpA* transcript is a prototypical example with its 35 nonsilent (and expected) sites edited to levels between 75% and 98% and the only silent site (atpAeU1068PP) edited to only 10% (Table S1). Similarly, in *psbC*, 15 expected nonsilent sites are edited to 91–99%, whereas an unexpected nonsilent 'extra' edit that does not reconstitute a conserved amino acid (psbCeC734IT) and two others in the 5′-UTR are edited only 5–11% (Table S1). However, few notable exceptions exist. The *psaB* gene features two silent sites (psaBeU15FF and psaBeU525LL) edited at high frequencies above 97% (Table S1). By contrast, removal of some stop codons is surprisingly inefficient; for example, 9% at the cysTeC163*Q site or only 6% at the ndhCeC175*Q site. Because these values were barely above our general criteria to reliably identify editing sites in the RNA-seq data (minimum 5% change at minimum 30-fold coverage), we rechecked similar positions with initially undetected edits, confirming that expected edits in *rpoB*, *rpoC2*, *ycf2*, and, most notably, in *ycf1* indeed exist but fell below those initial quality thresholds (Table S1).

We verified numerous predicted RNA editing sites in *chlL*, *ccsA*, *ndhE*, *petN*, *psaJ*, *psbN*, *rpl14*, and *rpl22* in *A. agrestis* that also seem to be necessary in the sister species *A. angustus*, but were likely missed in an early RT-PCR-based editing study (Kugita *et al.*, 2003b). Altogether, the nonsilent edits identified in coding sequences match the predictions very well. Only 21 strongly predicted editing events remained unconfirmed, 12 nonsilent edits were unexpected, and 52 fell below the initial prediction threshold owing to lack of amino acid sequence conservation at these sites (Table S1).

A particularly interesting finding concerns the divergent RNA editing patterns in the two *Anthoceros* sister species (Fig. 2). An edited nucleotide in one species may be 'pre-edited' with the appropriate pyrimidine already present in the cpDNA of the respective other species. Strikingly, the majority of C-to-U edits are shared between the two taxa, whereas reverse U-to-C editing sites are much more variable, most often with *A. agrestis* requiring U-to-C editing where *A. angustus* features a cytidine at genomic level (Fig. 2a). The *psbC* and *psbD* genes are prominent examples (Fig. 2b,c). Of 44 nonsilent edits – 17 in *psbC* and 27 in *psbD* – only 12 are shared between the two *Anthoceros* species, whereas 32 occur exclusively in one taxon to reconstitute on a transcript level what is genomically present in the other. Most importantly, 26 of these unique editing sites in these two genes (i.e. 80%) are U-to-C edits in *A. agrestis*.

RNA editing contributes to classifying organellar genes as functional or as pseudogenes, most notably given the numerous necessary conversions of stops into glutamine or arginine codons through U-to-C editing. An intriguing case is *matK*, the maturase in the group II intron of the *trnK* gene, highly conserved among plants. We confirmed 11 edits as predicted, but only at very low frequencies (Table S1). More importantly, we could not identify editing at 12 further sites including necessary removals of eight stop codons, even after rechecking whether they had been missed owing to the initial threshold levels. Hence, we consider *matK* a pseudogene also in *A. agrestis*, yet in an earlier state of degeneration than its counterpart in *A. angustus* (Kugita *et al.*, 2003b). Notably, however, the host gene of the corresponding intron, *trnK*, is subject to efficient editing as expected. As in *A. angustus*, the *trnK* gene features a UUC anticodon at the DNA level that would erroneously match GAR glutamate codons, but which is edited into a UUU anticodon to properly match AAR lysine (K) codons instead.

Whereas RNA editing patterns suggest *matK* to be a pseudogene, exactly the opposite is observed for the hornwort-specific small *orf51* downstream of *rps16*, lacking sequence similarities to other proteins. Here, we find that three RNA editing sites are shared with *A. angustus*, supporting a possible functional role.

The *Anthoceros agrestis* chondrome

The *A. agrestis* mtDNA of 227 925 bp (Fig. 3) is larger than those in other hornworts previously investigated: 184 908 bp in *N. aenigmaticus*, earlier designated as *Megaceros aenigmaticus* (Li *et al.*, 2009), 209 482 bp in *Phaeoceros laevis* (Xue *et al.*, 2010) or 212 153 bp in *L. dussii* (Villarreal Aguilar *et al.*, 2018). However, it is surpassed in size by the very recently determined mtDNA sequence of 242 410 bp in the sister taxon *A. angustus* (Dong *et al.*, 2018). The extended *Anthoceros* chondrome sizes result from a larger set of introns, larger intron sequences, less pseudogene degeneration, and larger intergenic sequences (IGSs; Fig. 3).

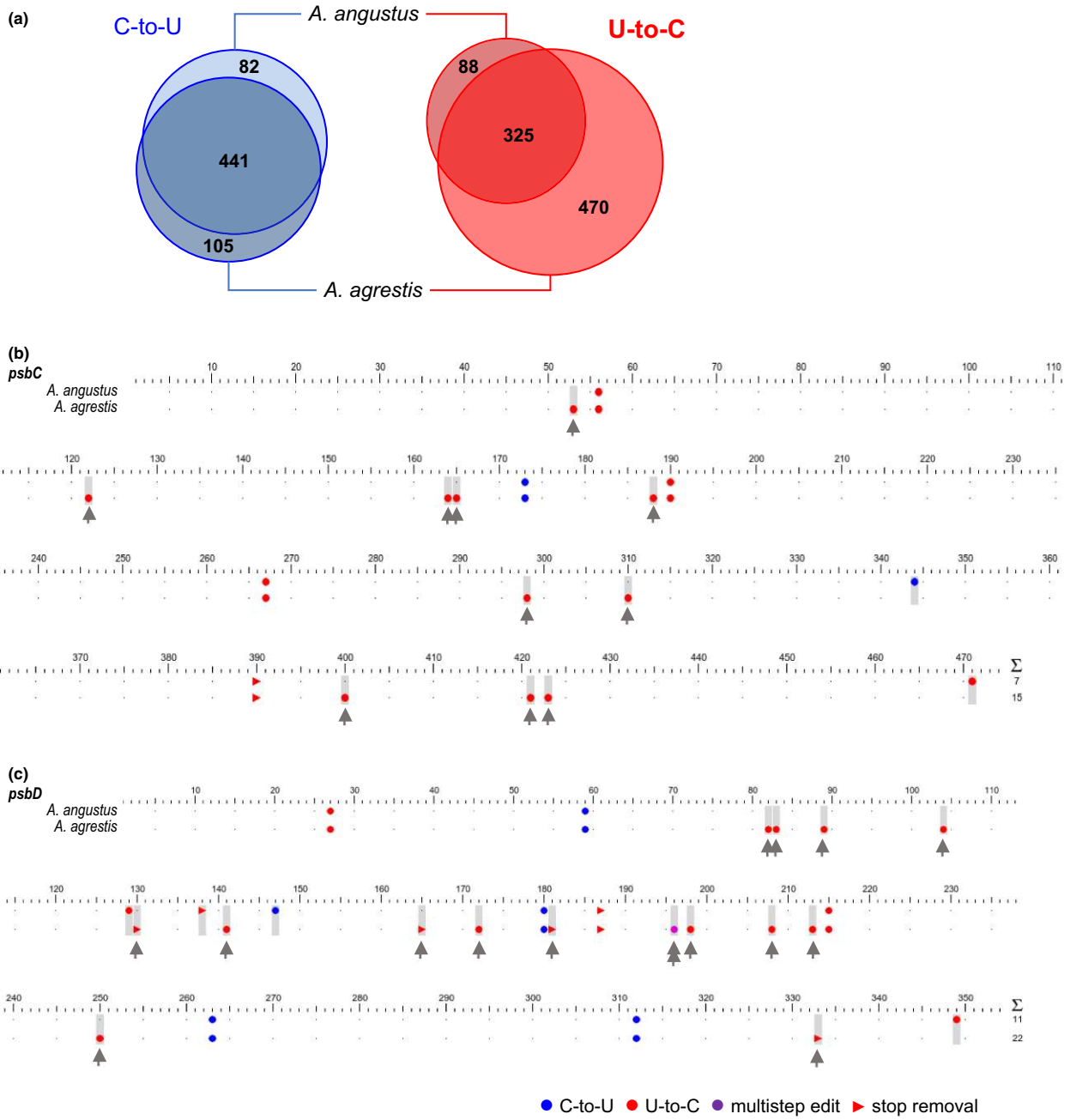


Fig. 2 (a) Venn diagrams summarizing chloroplast nonsilent RNA editing events in coding regions in *Anthoceros agrestis* (this study) and *Anthoceros angustus* (Kugita *et al.*, 2003b). Most C-to-U editing sites (441) are shared between the two sister taxa, whereas most U-to-C editing sites (470) are exclusively present in *A. agrestis*. (b, c) The chloroplast *psbC* (b) and *psbD* (c) genes are given as examples for the differences in the editing patterns. Editing overview panels have been created with PREPACT (Lenz *et al.*, 2018). Scales on top indicate codon numbering; a symbol legend is shown at the bottom. Only 12 of altogether 44 editing sites are shared, whereas 32 are species specific (grey shading) to reconstitute codons conserved at the genomic level in the other species. Among the species-specific edits, 26 are U-to-C editing events exclusively occurring in *A. agrestis* (grey arrows). Codon 196 in *psbD* is affected by multistep editing psbDeCC586LP (twice U-to-C, purple dot) to convert a leucine into a proline codon (double arrowhead).

The mitochondrial gene complement

The *A. agrestis* mtDNA reflects an evolutionarily ancient state with intact genes (*atp8*, *rpl2* and *trnS-GCU*) that are degenerated or missing in other hornworts (Table 2). We explicitly include here the analysis of intron-encoded maturases in group II introns, which have been somewhat neglected in the previous studies and adapt a systematic maturase nomenclature as suggested (Guo & Mower, 2013). Maturase loci are labelled with *mat* followed by a hyphen and the systematic name for the respective host intron. Accordingly, the *matR* locus, conserved in other plants (but degenerated into a pseudogene in hornworts), would become *mat-nad1i728g2*. The *A. agrestis* chondrome carries two, most likely functional, maturases in other introns: *mat-cox2i373g2* and

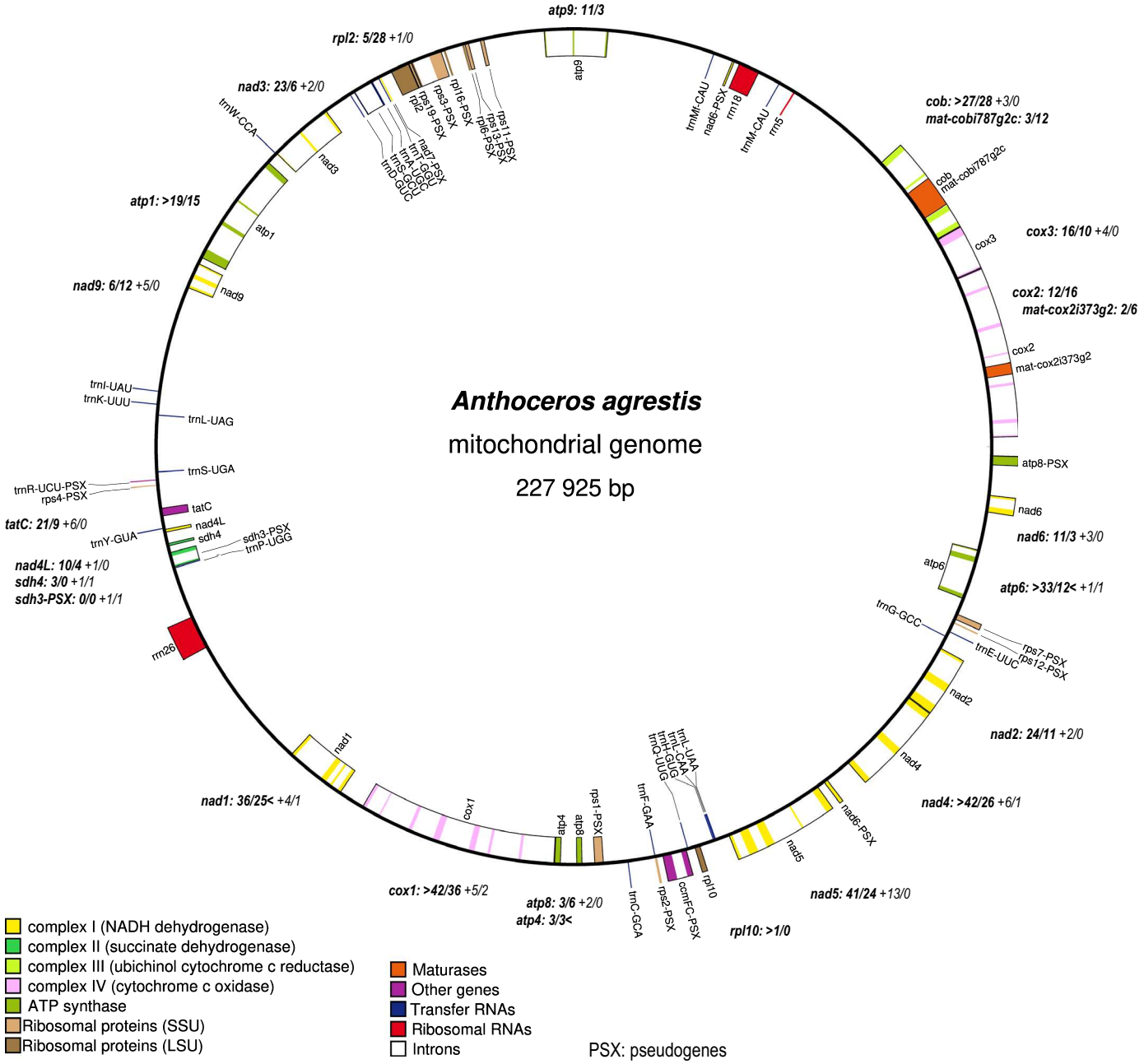


Fig. 3 The *Anthoceros agrestis* mitochondrial genome. As in other bryophytes, the chondrome maps as a circular molecule. The most notable differences to the mitochondrial DNAs (mtDNAs) of other hornworts concern less degenerated pseudogenes (PSX), more recognizable pseudogene traces, and apparently intact genes degenerated in the other species. The *A. agrestis* mtDNA is deposited in the database under accession no. MK087647. The genome map was drawn with OGDRAW (Lohse *et al.*, 2013). Gene categories are indicated in the legend. The display of RNA editing sites is like in Fig. 1.

mat-cobi787g2c (Fig. 3). We suggest an added 'c' in the latter case to indicate that the maturase ORF is continuous with the upstream reading frame of the host gene. Like its *cob* host gene, *mat-cobi787g2c* is heavily edited, including several sites evidently reconstituting amino acid positions conserved among maturases (Table S2).

Mitochondrial introns

The mtDNA of *A. agrestis* sets a record for an embryophyte organelle genome with altogether 44 introns, several of which are absent in the other hornwort genera (Table 2). Four of the *A. agrestis* mitochondrial introns are of group I (g1) and 40 are of the group II (g2) type, the latter including clearly detectable 'fossil' introns in degenerated pseudogenes *ccmFC*, *rps3* and *sdh3*. All of these are conserved in the very recently determined mtDNA of *A. angustus* (Dong *et al.*, 2018), for which we suggest a reinterpretation of some gene structures.

Intron atp9i87g2 is conserved in liverworts, mosses, and lycophytes but is absent in *Nothoceros* and *Phaeoceros*. Likely owing to the tiny second *atp9* exon of only 8 bp, it has been missed in the annotation of the *A. angustus* mtDNA where it was erroneously

Table 2 The *Anthoceros agrestis* mitochondrial DNA gene and intron complement compared with those of *Anthoceros angustus* (Dong *et al.*, 2018), *Nothoceros aenigmaticus* (Li *et al.*, 2009), and *Phaeoceros laevis* (Xue *et al.*, 2010).

	A. agrestis	*A. angustus*	*N. aenigmaticus*	*P. laevis*		*A. agrestis*	*A. angustus*	*N. aenigmaticus*	*P. laevis*		*A. agrestis*	*A. angustus*	*N. aenigmaticus*	*P. laevis*
atp1	+	+	+	+	***nad1***	+	+	+	+	***rps12***	ψ	ψ	ψ	ψ
atp1i805g2	+	+	+	+	nad1i287g2	+	+	+	+	***rps13***	ψ	ψ	+	+
atp1i1019g2	+	+	+	+	nad1i348g2	+	+	+	+	***rps14***			+	
atp1i1050g2	+	+	+	+	nad1i728g2M	+	+	+	+	***rps19***	ψ	ψ		
atp4	+	+	+	+	***nad2***	+	+	+	+	***rrn5***	+	+	+	+
atp6	+	+	+	+	nad2i709g2	+	+	+	+	***rrn26***	+	+	+	+
atp6i80g2	+	+	+	+	nad2i1282g2	+	+	+	+	***rrn18***	+	+	+	+
atp6i439g2	+	+	+	+	***nad3***	+	+	+	+	***sdh3***	ψ	ψ	ψ	ψ
atp8	+	+	ψ	ψ	nad3i52g2	+	+	+	+	*sdh3i100g2*	+	+	+	+
atp9	+	+	+	+	nad3i140g2	+	+	+	+	***sdh4***	+	+	+	+
atp9i87g2	+	+			***nad4***	+	+	+	+	***tatC***	+	+	+	+
atp9i95g2	+	+	+	+	nad4i461g2	+	+	+	+					
ccmFC	ψ	ψ	ψ	ψ	nad4i976g2	+	+	+	+	***trnA-UGC***	+	+	+	+
ccmFCi829g2	+	+	+	+	***nad4L***	+	+	+	+	***trnC-GCA***	+	+	+	+
cob	+	+	+	+	***nad5***	+	+	+	+	***trnD-GUC***	+	+	+	+
cobi420g1	+	+		+	nad5i230g2	+	+	+	+	***trnE-UUC***	+	+	+	+
cobi787g2Mc	+	+	+	+	nad5i881g2	+	+			***trnF-GAA***	+	+	+	+
cobi838g2	+	+	+	+	nad5i1455g2	+	+	+	+	***trnG-GCC***	+	+	+	+
cox1	+	+	+	+	nad5i1477g2	+	+	+	+	***trnG-UCC***	ψ	ψ	ψ	ψ
cox1i44g2	+	+	+	+	***nad6***	+	+	+	+	***trnH-GUG***	+	+	+	+
cox1i150g2	+	+	+	+	nad6i444g2	+	+	+	+	***trnI-CAU***	+	+	+	+
cox1i253g1	+	+			***nad7***	ψ	ψ	ψ	ψ	***trnK-UUU***	+	+	+	+
cox1i653g2	+	+			***nad9***	+	+	+	+	***trnL-CAA***	+	+	+	+
cox1i1116g1	+	+			nad9i246g2	+	+	+	+	***trnL-UAA***	+	+	+	+
cox1i1298g2	+	+	+	+	nad9i502g2	+	+	+	+	***trnL-UAG***	+	+	ψ	+
cox1i1305g1	+	+	+	+	***rpl2***	+	+			***trnM-CAU***	+	+	+	+
cox2	+	+	+	+	***rpl5***			ψ		***trnMf-CAU***	+	+	+	+
cox2i98g2	+	+			***rpl6***	ψ	ψ	ψ	ψ	***trnP-UGG***	+	+	+	+
cox2i281g2	+	+	+	+	***rpl10***	+	+	+	+	***trnQ-UUG***	+	+	+	+
cox2i373g2M	+	+	+	+	***rpl16***	ψ	ψ			***trnR-UCU***	ψ	ψ	ψ	ψ
cox2i381g2	+	+		+	***rps1***	ψ	ψ	ψ	ψ	***trnS-GCU***	+	+	ψ	ψ
cox2i564g2	+	+		+	***rps2***	ψ	ψ		ψ	trnS-GCUi43g2	+	+	ψ	ψ
cox2i691g2	+	+		+	***rps3***	ψ	ψ			***trnS-UGA***	+	+	ψ	ψ
cox3	+	+	+	+	rps3i74g2	+	+			***trnT-GGU***	+	+	+	+
cox3i109g2	+	+		+	***rps4***	ψ	ψ	ψ	ψ	***trnV-UAC***	ψ	ψ	ψ	ψ
mat-cobi787g2c	+	+	+	ψ	***rps7***	ψ	ψ	ψ	ψ	***trnW-CCA***	+	+	+	+
mat-cox2i373g2	+	+	ψ	+	***rps8***			ψ		***trnY-GUA***	+	+	+	+
mat-nad1i728g2	ψ	ψ	ψ	ψ	***rps11***	ψ	ψ	ψ	ψ					

Our assessment on gene and intron complements differs from those reported in the previous studies in several instances, as exemplarily discussed in the main text for selected examples. Pseudogenes are indicated by ψ. Yellow shading highlights group II (g2), orange shading indicates group I (g1) introns. Introns are labelled as previously suggested (Dombrovska & Qiu, 2004; Knoop, 2004), and a nomenclature proposal (Guo & Mower, 2013) is adapted to consistently label intron-encoded maturases ('mat') to properly indicate their respective host introns (see main text).

merged with atp9i95g2 (Table 2). *Anthoceros* features three additional introns in *cox1*: cox1i249g1, cox1i653g2, and cox1i1116g1. Introns cox1i249g1 and cox1i653g2 at present share no significant similarities with any other sequence in the databases. We also detected the terminal group I intron cox1i1305g1 that has been overlooked in previous hornwort mtDNA studies where – again owing to a tiny exon of only 7 bp – a larger upstream cox1i1298g2 was erroneously annotated previously (Table 2). Intron nad5i881g2, initially detected in *Anthoceros punctatus* (Beckert *et al.*, 1999), is conserved in the other *Anthoceros* species and in *Leiosporoceros* but absent in the other hornworts.

Most interestingly, we could identify rps3i74g2 in the *A. agrestis rps3* pseudogene (Table 2), an intron previously detected only in vascular plants. Not listed in the earlier surveys (Dong *et al.*, 2018), we now find that *rps3* (including rps3i74g2) and *rpl16* are present as pseudogenes in the mtDNA of *A. angustus*, too. The discovery of rps3i74g2 adds to the candidate synapomorphies of an HT clade. Finally, we identified the *trnS-GCU* gene including intron trnS-GCUi43g2, conserved in liverworts and overlooked in the previous hornwort mtDNA annotations (Fig. 3; Table 2).

The *Anthoceros agrestis* mitochondrial editome

We identified 496 events of C-to-U and 403 sites of U-to-C editing in the mitochondrial transcriptome (Table S2). Hence, chloroplast exceeds mitochondrial RNA editing both in total numbers (1549 vs 899) and in the U-to-C/C-to-U ratio (58% vs 46%). As in the chloroplast, mitochondrial RNA editing mostly reconstitutes codon identities as predicted (Table S2). Among the rare edits in IGSs, four were found in the large and pseudogene-rich IGS between *rpl2* and *atp9*, possibly as a leftover of formerly functional *rps3* and *rpl16* editing, and 10 edits were identified in the pseudogene-rich IGS between *atp8* and *nad5*, not affecting the (likely) *rpl10* pseudogene, however (Fig. 3).

Like in the chloroplast editome, we observed surprisingly inefficient stop codon removal in a few cases (e.g. of only 20% for cox2eC55*Q). Most importantly, RNA editing efficiencies consistently remained low for the *tatC* gene also in independent RT-PCR approaches, and we found only marginal evidence (< 3%) for the necessary stop codon conversion tatCeC511*Q (Table S2), leaving the status of *tatC* as a functional gene dubious. Editing in the other mitochondrial genes, however, is largely as predicted. We use the *cox1* gene here as an example (Fig. 4), and also later in the interest of discussing candidate site-specific editing factors.

Mitochondrial RNA editing in 5′ and 3′-UTRs and in other structural RNAs is low (Table S2). One noteworthy exception is the tRNA-Asp (GUC) encoded by the *trnD-GUC* gene. Two events of U-to-C editing strengthen base pairing in the dihydrouridine and in the pseudouridine stem by converting G–U into G–C base pairs, but a further candidate C-to-U edit to create an additional base pair in the anticodon stem was not detected (Fig. S1).

Strikingly, we detected many edits in mitochondrial introns. A surprising number of 61 RNA editing sites in both directions of pyrimidine exchange cluster in the characteristic domains V and VI at the end of group II introns (Fig. 5). Most of these editing events in *A. agrestis* contribute to stabilizing the characteristic structures of these two small domains at the 3′ intron ends and exist in a 'pre-edited' state at the mtDNA level in the homologous introns of other hornworts. Notably, chloroplast introns were much less affected despite an overall dominance of chloroplast over mitochondrial RNA editing (Table S1). One U-to-C editing event in consensus position 29 of domain V is shared by 19 group II introns (Figs 5, 10; see later). Later, we discuss a candidate RNA editing factor with an excellent match to the domain V sequences conserved in seven of these introns (Fig. 10; see later).

The diverse *Anthoceros agrestis* nuclear PPR gene family

The recently redefined HMMER profiles for the different types of PPRs and for the E1, E2 and DYW domains (Cheng *et al.*, 2016) were used to scan the *A. agrestis* genome assemblies. We identified a total of 3089 loci encoding PPR proteins. Of these, only 145 were of the P-type containing only canonical P-type PPRs, whereas most protein models featured PLS-type PPRs, as typical for hitherto identified C-to-U RNA editing factors. Among the latter, 1480 were initially scored as 'pure' PLS-type PPR proteins lacking recognizable additional carboxyterminal domains, 77 with an E1 domain, 447 with an E1 and E2 domain, and 928 as E+ proteins (i.e. C-terminally extended beyond their E2 domain). Only 12 protein models were initially classified as having a canonical DYW domain, including three with no extensive upstream PPR array.

Carefully reinspecting the initially identified loci revealed that many of these in fact feature highly deviant DYW domain variants and/or DYW domain truncations. A phylogenetic tree of the *A. agrestis* PLS-type proteins reassessed as extending beyond an E2 domain – hence including canonical, divergent, and truncated DYW domain variants – is shown in Fig. 6. All nine RNA editing factors of *P. patens* and 28 *A. thaliana* RNA editing factors with full-length DYW domains were used as an outgroup.

The extended DYW protein family *sensu lato* in *A. agrestis* falls into distinct clades (Fig. 6) featuring significant deviations from the DYW domains in the hitherto identified C-to-U editing factors (Fig. 7). Only a few *A. agrestis* proteins have a complete DYW domain with a conserved N-terminal 'PG box' including the characteristic PGxSWIE motif (Okuda *et al.*, 2007; Hayes *et al.*, 2013). They are accompanied by clades including C-terminally truncated and many 'WW-type' homologues that feature a notable variant of the PG box with the tryptophan (W) in position 5 of the PGxSWIE motif duplicated (Figs 6, 7). Most proteins (734) of the *A. agrestis* gene family, however, are placed in a superclade of proteins with generally full-length, but highly diverged DYW domains with characteristic differences in conserved amino acid positions (Fig. 7). Proteins of this superclade are characterized by a significantly different PG box ('KPAxA') and fall into two subtypes with 'DRH' or 'GRP' replacing the eponymous DYW tripeptide at the end of the 'classic' DYW domain.

Within the KPAxA superclade, the GRP-type proteins dominate in number and seem to be a more recent expansion of the gene family emerging from the DRH-type proteins (Fig. 6).

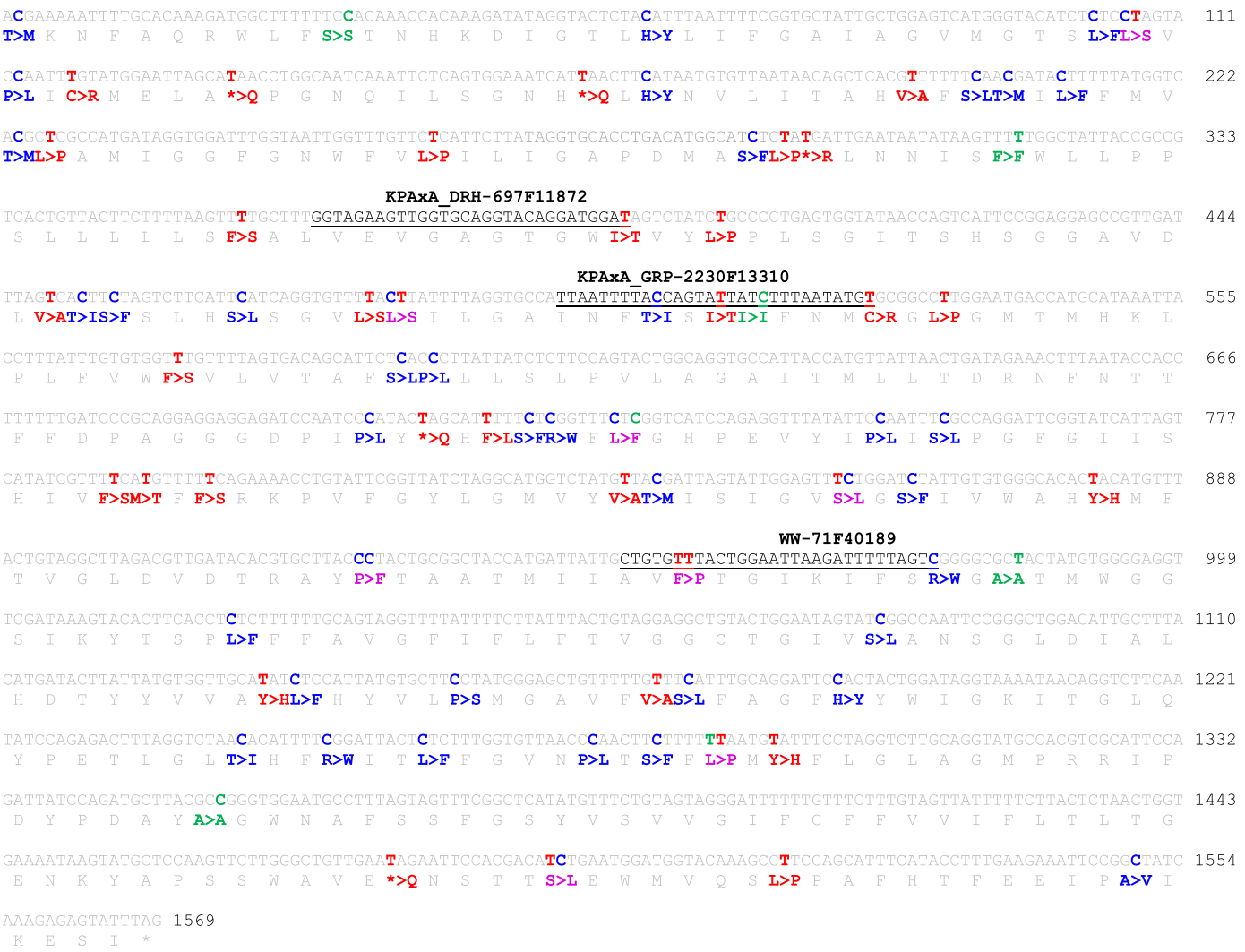


Fig. 4 The *cox1* gene as an example for mitochondrial RNA editing in *Anthoceros agrestis*. The *cox1* exons show 80 sites of C-to-U (blue) and U-to-C (red) editing, predictably reconstituting conserved codons. Silent editing sites are shown in green font and codons affected by multiple editing in purple. The output is based on the complementary DNA analysis function in PREPACT (Lenz *et al.*, 2018). Some 50% of *A. agrestis* edits are shared with lycophytes *Isoetes engelmannii* and/or *Selaginella moellendorffii* included in the PREPACT 3.0 editome references. RNA editing of the *cox1* start codon simultaneously creates the stop codon for *atp4*, which overlaps by 4 bp. Top-scoring candidate binding regions are indicated exemplarily for PLS-type PPR proteins of the DRH, GRP and WW types now identified in *A. agrestis*, which show characteristic differences to canonical DYW domains (see Figs 6, 7).

Other than a joint four-amino-acid deletion, many differences in amino acid conservation (most notably G3A, S5A, D21E, Y47H, S71A, I76L, I89L, V97M, K110R, F124V, and Y137R) are shared between all KPAxA-type DYW domains (Fig. 7). Additional differences (most notably V4, T6, A31, L32, V34, T37, P57, S85, V90, E106, D115, G121, Y122, A128, K135, G136 and P138) are seen in GRP-type DYW domains alone (Fig. 7).

Candidate factors for C-to-U and U-to-C RNA editing

Previously characterized C-to-U RNA editing factors are PLS-type proteins with a complete canonical DYW domain or, when truncated, extend across a recognizable PG box motif and acquire the DYW cytidine deaminase activity *in trans*. Hence, we consider most of the DYW variant proteins in *A. agrestis*, including the truncated versions with recognizable PG boxes and their variants, prime editing factor candidates, also considering their large number correlating well with the abundant RNA editing now identified. The high number of proteins with deviant DYW domains is intriguing in the light of the high amount of reverse U-to-C RNA editing that we could identify. Naturally, the characteristic DRH and GRP DYW domain variants (Fig. 7) could be attractive candidates to represent factors for reverse U-to-C RNA editing.

We used the consensus motifs for the different DYW protein clades now identified in *A. agrestis* (Fig. 7) as queries to scan available genome and transcriptome data (GenBank/NCBI and OneKP data). We could identify DYW proteins with a canonical PG box in all land plant clades except the marchantiid (complex-thalloid) liverworts, lacking RNA editing altogether and fitting the previous surveys on RNA editing (Rüdinger *et al.*, 2012). By contrast, KPAxA-type DYW domains were exclusively

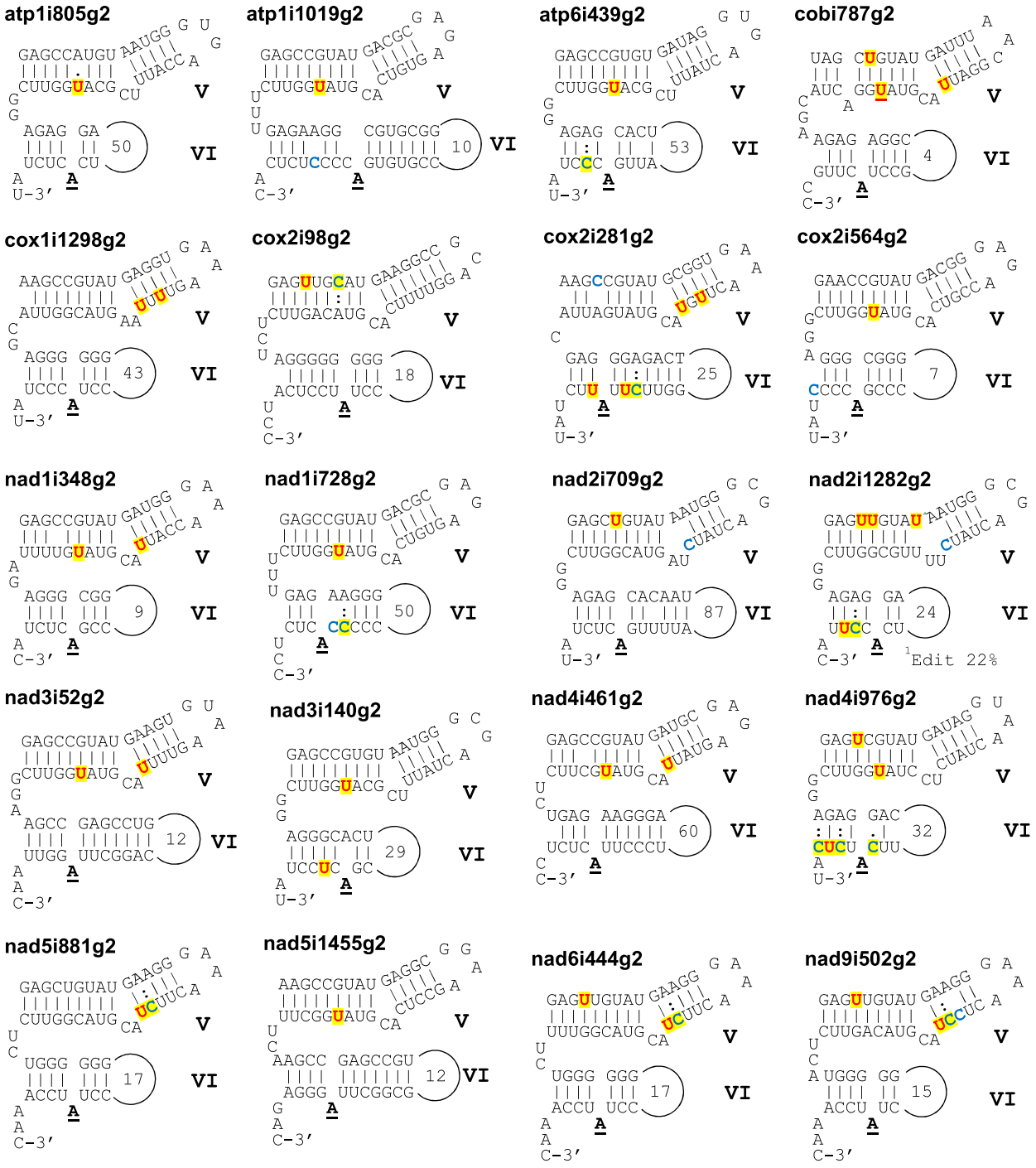


Fig. 5 Numerous events of RNA editing (yellow background) were identified in terminal domains V and VI of mitochondrial group II introns and contribute to stabilizing their canonical secondary structures by converting G–U into G–C pairs (red) or establishing A–U pairs from A–C mismatches (blue, colons). The bulged A for lariat formation in domain VI is highlighted (bold, underlined). Two cases of apparent misediting weakening base pairs in atp1i805g2 and nad4i976g2 occur at low frequencies only (Supporting Information Table S2). Seven additional C-to-U edits could be expected for atp1i1019g2, cox2i281g2, cox2i564g2, nad1i728g2, nad2i709g2, nad2i1282g2 and nad9i502g2 (blue font, no background) but were not observed in the transcriptome data. Noteworthy is the U-to-C editing event in domain V consensus position 29 (underlined in the cobi787g2 example, top right) occurring in altogether 19 introns. This editing event is also identified in seven further introns (Fig. 10) where domain V sequences match a candidate RNA editing factor.

identified in the available transcriptome data of hornworts, ferns, and lycophytes. We could not find evidence for KPAxA-type proteins in mosses, liverworts, seed plants, or in the Selaginellales, where, despite most extensive C-to-U editing, not a single case of reverse U-to-C editing had been identified (Hecht *et al.*, 2011; Oldenkott *et al.*, 2014). Hence, the presence of the now discovered KPAxA-type DYW domains with their divergent amino acid signatures (Fig. 7) seems to have a perfect

New Phytologist

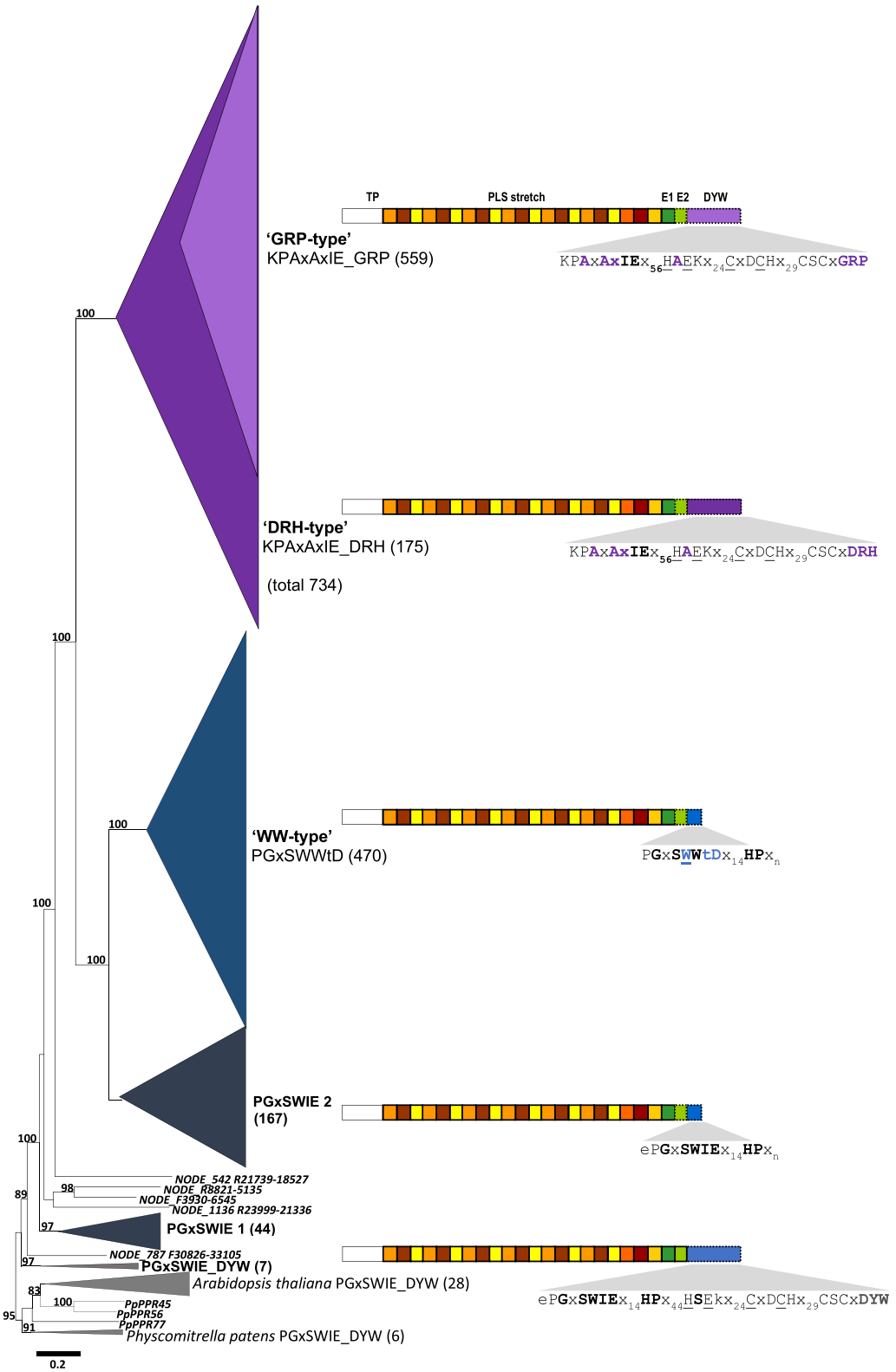


Fig. 6 Maximum likelihood phylogenetic tree of 1428 PLS-type PPR proteins identified in the nuclear genome assembly of *Anthoceros agrestis*, which carry carboxyterminal domains extending into recognizable full or truncated DYW domains. The DYW-type PPR proteins of *Physcomitrella patens* and those identified as C-to-U RNA editing factors in *Arabidopsis thaliana* were used to root the gene family tree. Only five proteins in *A. agrestis* have full-length canonical DYW domains; many others are variably truncated behind the 'PG box'. Most proteins in *A. agrestis* with characteristic alterations in their DYW domain signatures ('WW-type', 'DRH-type' and 'GRP-type'; see also Fig. 7) fall into clades with significant bootstrap support. Protein models are shown next to each clade, with significant amino acid changes indicated in the colour of the respective collapsed clade. Amino acids of the conserved cytidine deaminase signature are underlined. Dotted lines indicate variable E2/DYW domain truncations in some members of the respective clades.

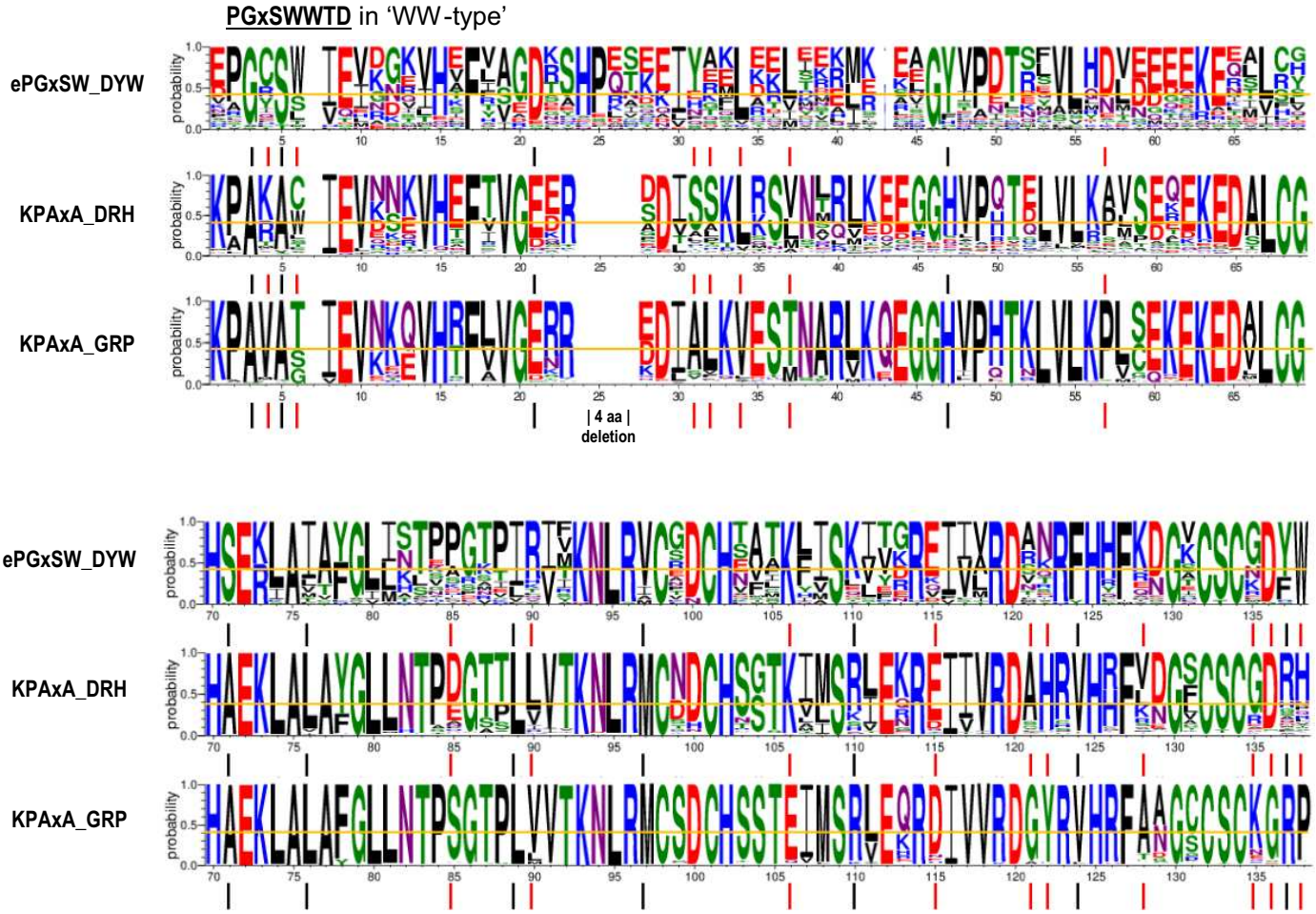


Fig. 7 Different sequence conservation profiles in the DYW domain variants 'DRH' and 'GRP' identified in *Anthoceros agrestis*. Conservation plots were created using the WebLogo service at http://weblogo.berkeley.edu/logo.cgi (Crooks *et al.*, 2004). The profile for canonical DYW domains (ePGxSW_DYW-type) is based on the alignment of proven C-to-U RNA editing factors with full-length DYW domains characterized in *Arabidopsis thaliana* (28 sequences) and *Physcomitrella patens* (nine sequences) and their five full-length homologues in *A. agrestis*. No significant differences in the conservation profile are observed for the 236 C-terminally truncated homologues in *A. agrestis* in the respective amino-terminal regions. Numerous characteristic changes in conserved positions along the entire DYW domain are observed among the KPAxA_DRH (168 sequences) and the KPAxA_GRP-type DYW proteins (482 sequences) identified in *A. agrestis*. Truncated DRH-type and GRP-type proteins lacking a DYW domain were excluded for the WebLogo creation. Significant changes in amino acids conserved at a threshold of at least 0.6 (orange lines) are highlighted with black lines for positions shared among all KPAxA-type DYW domains and with red lines for those in the GRP-type proteins alone. Other than shifts in amino acid conservation, the KPAxA-type proteins share a deletion of four amino acids (alignment positions 24–27). The amino-terminal PGSWIE heptapeptide motif of the 'PG box' in the canonical DYW domain is changed to PGSWWTD including a tryptophan (W) duplication in alignment position 6 in 474 truncated 'WW-type' proteins (top left). No differences are seen, however, in the highly conserved motifs $H^{70}xEx_nCxxC^{101}$ and $H^{125}xFx_4CSC^{134}$, which are very likely relevant for binding zinc ions (Hayes *et al.*, 2013).

phylogenetic overlap with taxa showing reverse U-to-C RNA editing.

Consequently, we strived to identify candidate targets in the organelle transcriptomes for the PPR arrays in front of the different DYW variant proteins. To this end, we used all reassessed proteins featuring at least a recognizable PG box variant and minimally 14 upstream PLS-type PPRs to extract possible RNA targeting information from positions 5 and L of their P and S-type PPRs (see Materials and Methods section). To avoid bias, no preselection for possible organelle targeting preference to chloroplasts or mitochondria was done, and top matches were scored both for searches of PPR arrays against candidate editing targets (Fig. 8a) and the other way around (Fig. 8b). This analysis revealed that the reverse U-to-C editing sites strongly dominate among the top candidate targets for the KPAxA-type proteins (Fig. 8). By contrast, the PPR arrays upstream of the classic PGxSWIE-type and of the WW-type variants preferentially matched sequences upstream of the now identified C-to-U editing positions.

Examples of top matches for one member each of the deviant WW, DRH, and GRP-type DYW proteins within the *cox1* mRNA are shown in Fig. 9. The PPR arrays in each of the proteins have at least 13 perfect matches to their potential targets upstream of the respective editing sites, with the WW-type protein matching to a C-to-U editing site and the DRH and GRP proteins potentially binding upstream of U-to-C editing sites.

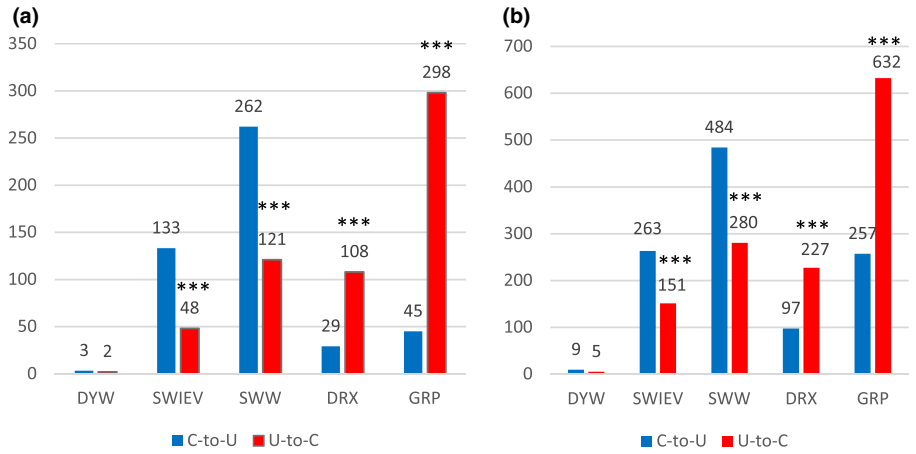


Fig. 8 Bar charts summarizing the respective top matches between members of the five different clades of DYW proteins now identified in *Anthoceros agrestis* (*x*-axis) and all C-to-U (blue) or U-to-C (red) editing sites now determined in the two organelles. Positions 5 and L were extracted from P and S-type PPRs of 1049 'DYW' proteins featuring at least 14 PPRs and translated into a scoring matrix following the PPR-RNA code rules (see the Materials and Methods section). (a) The numbers of respective top matching C-to-U or U-to-C editing sites in the complete *A. agrestis* organelle editome for the members of the five different DYW protein clades. The respective numbers of PPR proteins per type are given above the bars. Only cases where the score of the best-fitting editing site is higher than the second-best hit are included. (b) Reciprocal assignments for the superset of 2405 editing sites identified among the top candidate targets for the different DYW-type proteins under A. ***, $P > 0.99$ for the preferred mutual assignments of DRX and GRP proteins to U-to-C edits and SWIEV and SWW-type DYW proteins to C-to-U edits over equal distributions (50% each) in the one-proportion *Z*-test.

We noted that among the RNA editing sites within group II intron domains V and VI, one event of U-to-C editing in domain V consensus position 29 is shared among 19 different mitochondrial group II introns (Figs 5, 10). Seven of these introns share extended similarities in their upstream domain V sequences. The PPR array of a GRP-type DYW protein matches excellently to the candidate target sequence upstream of this shared U-to-C editing site (Fig. 10). Intriguingly, matches are not only observed for the P and S-type PPRs according to the PPR-RNA recognition code rules but may be extended to two of the L-type repeats (L-5SN and L-8SN) potentially matching the conserved adenines in the corresponding targets, as recently observed for the moss DYW protein PPR65 targeting the *ccmFC* RNA (Oldenkott *et al.*, 2019).

Discussion

Despite being the species poorest of all major clades of extant land plants, hornworts are fundamentally important to understand the backbone phylogeny of embryophytes (Villarreal *et al.*, 2013). Fossils like *Horneophyton* may represent 'evolutionary bridges' between bryophytes and early tracheophytes (e.g. Hetherington & Dolan, 2018), but no morphological or developmental synapomorphies conclusively resolve the phylogeny of the three extant bryophyte clades relative to tracheophytes. The discussion has recently been reactivated with analyses of large nuclear transcriptome data sets (Wickett *et al.*, 2014; Cox, 2018; de Sousa *et al.*, 2018; Morris *et al.*, 2018; Puttick *et al.*, 2018; Rensing, 2018a,b), questioning the previously suggested phylogeny with liverworts sister to all other embryophytes and an HT clade (Qiu *et al.*, 1998; Groth-Malonek *et al.*, 2005; Qiu *et al.*, 2006). Notably, however, the latter phylogeny was identified again in a recent study using concatenated chloroplast genes with broad embryophyte taxon sampling (Lutzoni *et al.*, 2018). The biochemical composition of cell-wall xyloglucans (Peña *et al.*, 2008; Schultink *et al.*, 2014) or the 'fossil' group II intron rps3i74g2 in *A. agrestis* identified here may also support an HT clade, similar to the nad5i1477g2 intron (Groth-Malonek *et al.*, 2005) or, possibly, the evolutionary gain of 'reverse' U-to-C RNA editing in land plant organelles.

The 'reverse' type of U-to-C RNA editing in plant organelles is frequently referred to as 'occasional', suggesting it to comprise rare events accompanying the dominant and near-omnipresent C-to-U editing in plant chloroplasts and mitochondria. However, U-to-C editing is clearly abundantly present in hornworts and ferns and can even be the dominant direction of pyrimidine exchange RNA editing, as reported here and in earlier studies (Kugita *et al.*, 2003b; Guo *et al.*, 2015; Knie *et al.*, 2016).

Independent of their exact phylogenetic position, hornworts will likely represent the most ancient plant clade featuring reverse U-to-C RNA editing in their organelle genomes. Hence, we consider them the best *a priori* choice for future functional studies of U-to-C RNA editing, likely retaining the ancestral features of the 'reverse' editing biochemistry. A small genome size of only 84 Mbp – notably in contrast to the voluminous and polyploid genomes of most ferns – and the current progress on establishing it as a new plant model system (Szövényi *et al.*, 2015) make *A. agrestis* particularly attractive for studies of U-to-C RNA editing.

With the assembled *A. agrestis* chloroplast and mitochondrial genomes and their complete editomes now available in addition to nuclear genome assemblies, the hornwort allows the correlation of abundant RNA editing in both directions of pyrimidine exchange with potential specificity factors. Analysing the vastly extended and diversified family of DYW-type PPR proteins in *A. agrestis* revealed three highly derived variants (referred to here as WW, DRH, and GRP; see Figs 6, 7) of the likely ancestral DYW domain characteristic of RNA editing factors identified in C-to-U-only RNA editing models like *Arabidopsis* or

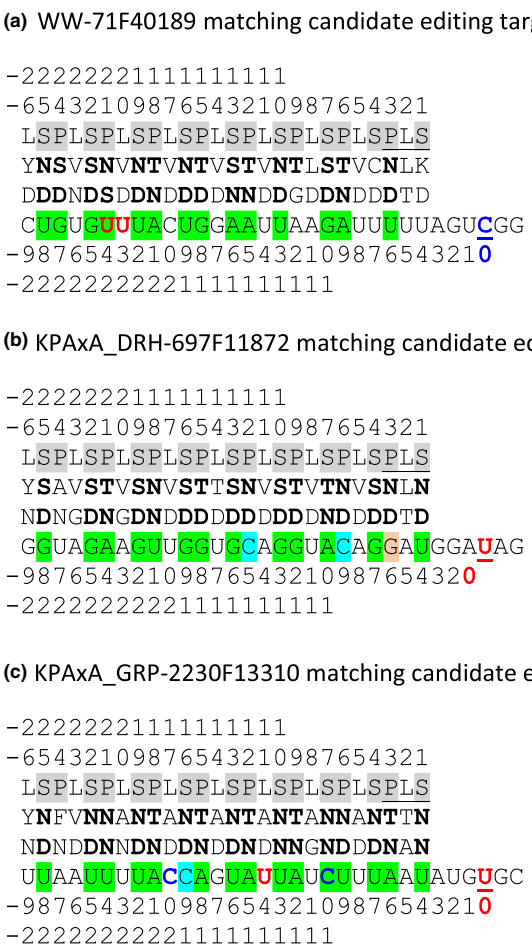


Fig. 9 Matches of selected PLS-type PPR proteins with noncanonical DYW domain variants (a) WW, (b) DRH, and (c) GRP having top-scoring candidate targets upstream of RNA editing sites in the *cox1* gene. The potential *cox1* targets (underlined sequences in Fig. 4) are the respective best-scoring sequences upstream of more than 2400 organelle editing sites now identified for each protein according to a scoring matrix following the PPR-RNA recognition rules (see the Materials and Methods section). Numbering runs backward both for the target sequence upstream of the editing sites and for the PPR arrays with the terminal S2-type PPR juxtaposed with target position −4. The terminal 'P2-L2-S2' PPR triplet with slightly differing amino acid signatures is underlined. Background shading indicates matches following the core RNA recognition rules for P and S-type PPRs (grey shading) according to amino acids in positions 5 and L (T/S+N: A; T/S+D: G; N+D: U; N+S: C; N+N: Y), with green indicating perfect matches, blue indicating pyrimidine transitions, and orange indicating mismatches. U-to-C editing is indicated in red; C-to-U editing is in blue.

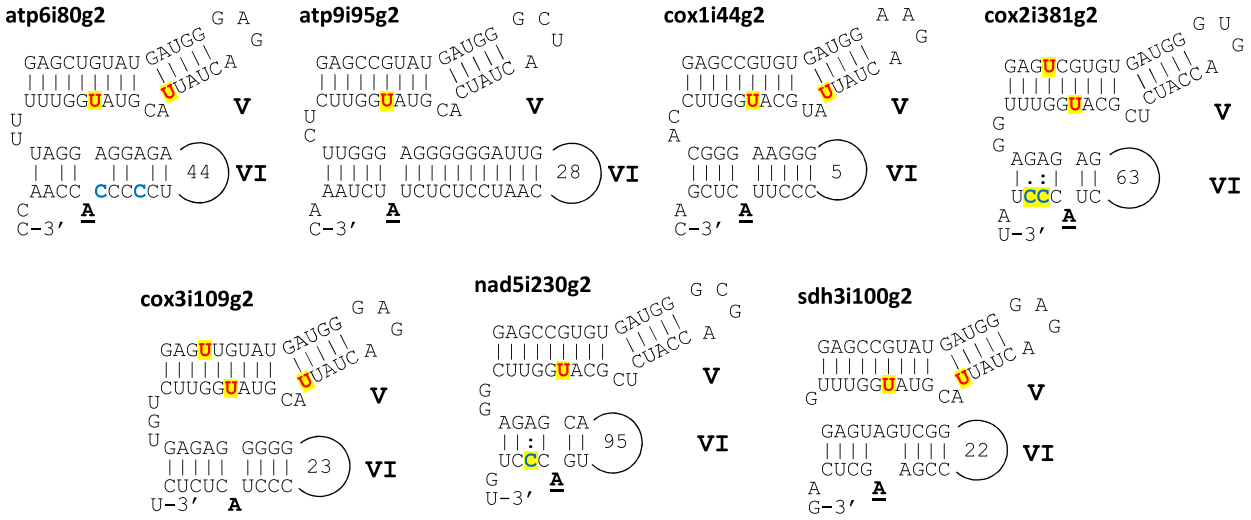
Physcomitrella. Intriguingly, we find that the DRH and GRP variants comprising the larger KPAxA clade now identified seem to have no homologues in plant taxa lacking U-to-C RNA editing and to match preferably to reverse U-to-C editing sites in *A. agrestis* (Figs 8–10). Additionally, the significant variability of U-to-C RNA editing in particular within *Anthoceros* (Fig. 2) may reveal valuable insights on the coevolution of editing sites and their cofactors.

One surprising additional result of our survey is the numerous edits in the small terminal domains V and VI of mitochondrial group II introns. The current understanding of target identification implies an alignment of PPRs and RNAs in a collinear fashion, but RNA secondary structures may interfere with this process. RNA editing may occur immediately after transcription before highly base-paired secondary structures are formed, or the binding of the PPR array may compete with secondary structure formation in an equilibrium of RNA molecules in different states.

Despite hundreds of conventional C-to-U edits along with the more abundant U-to-C edits, we found most of the 'classic' DYW domains typical of C-to-U editing factors to be C-terminally truncated in *A. agrestis* (Fig. 6). In the angiosperm models, such truncations are compensated for by separate DYW domains provided *in trans* (Boussardon *et al.*, 2012; Andrés-Colás *et al.*, 2017; Diaz *et al.*, 2017; Guillaumot *et al.*, 2017). Interestingly, we also identified three small DYW-only proteins outside of the large PLS-type PPR gene family in the *A. agrestis* genome, which feature the conserved cytidine deaminase signatures and a terminal DYW tripeptide (Fig. S2). Judged from transcript coverage, these three genes are more highly expressed than the PLS-type proteins, just as previously found for DYW2 in angiosperms (Andrés-Colás *et al.*, 2017). Hence, they could represent DYW domains to be supplied *in trans* for the many truncated proteins in *A. agrestis*, similar to the C-to-U editing setup in angiosperms. Disruption of single-polypeptide RNA editing factors like in the moss *Physcomitrella* (Schallenberg-Rüdinger & Knoop, 2016) may have occurred independently or may be yet another synapomorphy of the HT clade. We found no evidence in *A. agrestis*, however, for additional, non-PPR 'helper' components identified in angiosperms, such as MORFs/RIPs indicative of more complex editosomes (Bentolila *et al.*, 2012; Takenaka *et al.*, 2012; Zehrmann *et al.*, 2015; Bayer-Császár *et al.*, 2017; Haag *et al.*, 2017).

By contrast, complete full-length DYW domains dominate among the here defined KPAxA-type DYW proteins (Fig. 6). Despite the numerous deviations in their amino acid conservation profiles, the cytidine deaminase signature for Zn^{2+} coordination (Salone *et al.*, 2007; Boussardon *et al.*, 2014; Hayes *et al.*, 2015; Wagoner *et al.*, 2015; Ichinose & Sugita, 2018) is highly conserved among these proteins (Fig. 7). Consequently, and despite all the many differences compared with the more widespread

(a)



(b) Candidate matches for KPAxA-GRP-8056F3

```
   -111111111
   -876543210987654321
    LSPLSSPLSPLSPLSPLS
5:  YNSVNTTANTSNNSNNTN
L:  SDDNDDDDDDNNDNDNSN
    GUGAUGGGUGACCAUCUCGCAUGG   cox2i381g2eC2207
    GUGAUGGGCGACCAUCUCGCAUGG   nad5i230g2eC725
    AUGAUGGGAGACUAUUACGUAUGG   cox3i109g2eC2930
    AUGAUGGGAGACUAUUACGUAUGG   sdh3i100g2eC884
    AUGAUGGGAGACUAUUACGUAUGG   atp6i80g2eC557
    GUGAUGGAAGACUAUUAUGCAUGG   cox1i44g2eC2909
    AUGAUGGGCUACUAUCACGUAUGG   atp9i95g2eC2845
   -109876543210987654321 0
   -221111111111
```

Fig. 10 A candidate U-to-C RNA editing factor for a conserved editing event in *Anthoceros agrestis* mitochondrial group II introns. (a) Graphic display of group II intron domains V and VI is like in Fig. 5. The U-to-C RNA editing event in consensus position 29 of domain V is shared by altogether 19 group II introns (see Fig. 5 for 12 additional cases). (b) The PPR array of the variant DYW protein KPAxA-GRP-8056F3 matches to the domain V sequences of seven introns upstream of the editing site in consensus position 29. Numbering and shading are like in Fig. 9. PPR-13 is of the 'SS'-type (italics). Other than for the P and S-type repeats (light grey), matches are also observed for L-8SN and L-5SN (dark grey) juxtaposed with conserved adenosines in positions −11 and −8, respectively.

'classic' counterpart (Fig. 7), the KPAxA-type DYW domain would be unlikely to use a completely different mechanism of catalysis. As a working hypothesis, the characteristic differences between the KPAxA-type and classic DYW domains may result in acceptance of an amino-group donor as a co-substrate for uridine amination. We assume that future work on the organelle editomes and candidate editing factors in *Anthoceros* presented here will help to elucidate those and alternative hypotheses. Possibly, such reverse editing factors may in the future even prove to operate in heterologous systems, as recently shown for C-to-U RNA editing factors of *Physcomitrella*, conferring C-to-U editing in *Escherichia coli* and ultimately confirming the 'classic' DYW domain as cytidine deaminase acting on polyribonucleotides (Oldenkott *et al.*, 2019). Possibly, the new *E. coli* assay system could also prove to perform reverse U-to-C RNA editing and would subsequently allow the analysis of its biochemistry in detail. The matches between candidate reverse editing factors and their potential targets identified here are evidently a good starting point in that direction. The simple bacterial system will be more straightforward and superior in allowing the screening of many more candidate factors and targets than by genetic transformation of established plant models like *Arabidopsis* or *Physcomitrella*. In parallel, we will aim for the creation of knockout lines for candidate reverse editing factors in *A. agrestis* to elucidate their function.

Comparing plant lifestyles gives no reasonable clues as to why some plant lineages (like the Selaginellales) have lost reverse RNA editing altogether, may have never possessed it in the first place (possibly mosses and liverworts, depending on the ultimately true phylogeny of the bryophyte clades), or why U-to-C editing may even dominate over C-to-U editing in other lineages (Knie *et al.*, 2016). Based on the working hypotheses presented here, the experimental approaches outlined herein will hopefully help to answer that puzzling evolutionary question or, for example, also why RNA editing evolves so dramatically fast in at least some genera, like *Amaranthus* or *Silene* among the angiosperms (Sloan *et al.*, 2010; Hein *et al.*, 2019), *Selaginella* among the lycophytes (Smith, 2019), *Adiantum* among ferns (Zumkeller *et al.*, 2016), or, as also demonstrated here for U-to-C editing, in *Anthoceros* among the hornworts.

Acknowledgements

Work on RNA editing is supported by a grant of the German Research Foundation (DFG grant no. SCHA1952/2-1) to MSR. PS and AN are thankful for the financial support of the Swiss National Science Foundation (grants 160004 and 131726), the Georges and Antoine Claraz Foundation (Switzerland), the US National Science Foundation, the 'Forschungskredit', and the University Research Priority Program 'Evolution in Action' of the University of Zurich. AN was also supported by the Foundation of German Business (sdw).

Author contributions

PG assembled organelle genomes and did editome analyses. AN and PS isolated nucleic acids and conducted NGS and assemblies. IS, HL, PG, BG and RM designed bioinformatic pipelines to analyse PPR proteins and editing targets. MSR did phylogenetic analyses. VK and MSR designed the study and coordinated experimental efforts. VK wrote and edited the manuscript after critical input from the co-authors.

ORCID

Bernard Gutmann https://orcid.org/0000-0003-4657-0925
Volker Knoop https://orcid.org/0000-0002-8485-9423
Henning Lenz https://orcid.org/0000-0002-8080-0328
Mareike Schallenberg-Rüdinger https://orcid.org/0000-0002-6874-4722
Ian Small https://orcid.org/0000-0001-5300-1216
Péter Szövényi https://orcid.org/0000-0002-0324-4639

References

Andrés-Colás N, Zhu Q, Takenaka M, De Rybel B, Weijers D, Van Der Straeten D. 2017. Multiple PPR protein interactions are involved in the RNA editing system in *Arabidopsis* mitochondria and plastids. *Proceedings of the National Academy of Sciences, USA* **114**: 8883–8888.

Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I. 2012. A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genetics* **8**: e1002910.

Barkan A, Small I. 2014. Pentatricopeptide repeat proteins in plants. *Annual Review of Plant Biology* **65**: 415–442.

Bayer-Császár E, Haag S, Jörg A, Glass F, Härtel B, Obata T, Meyer EH, Brennicke A, Takenaka M. 2017. The conserved domain in MORF proteins has distinct affinities to the PPR and E elements in PPR RNA editing factors. *Biochimica et Biophysica Acta (BBA) – Gene Regulatory Mechanisms* **1860**: 813–828.

Beckert S, Steinhauser S, Muhle H, Knoop V. 1999. A molecular phylogeny of bryophytes based on nucleotide sequences of the mitochondrial *nad5* gene. *Plant Systematics and Evolution* **218**: 179–192.

Bentolila S, Heller WP, Sun T, Babina AM, Friso G, van Wijk KJ, Hanson MR. 2012. RIP1, a member of an *Arabidopsis* protein family, interacts with the protein RARE1 and broadly affects RNA editing. *Proceedings of the National Academy of Sciences, USA* **109**: E1453–E1661.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Boussardon C, Avon A, Kindgren P, Bond CS, Challenor M, Lurin C, Small I. 2014. The cytidine deaminase signature HxE(x)$_n$CxxC of DYW1 binds zinc and is necessary for RNA editing of *ndhD-1*. *New Phytologist* **203**: 1090–1095.

Boussardon C, Salone V, Avon A, Berthome R, Hammani K, Okuda K, Shikanai T, Small I, Lurin C. 2012. Two interacting proteins are necessary for the editing of the *ndhD-1* site in *Arabidopsis* plastids. *Plant Cell* **24**: 3684–3694.

Cheng S, Gutmann B, Zhong X, Ye Y, Fisher MF, Bai F, Castleden I, Song Y, Song B, Huang J *et al.* 2016. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *The Plant Journal* **85**: 532–547.

Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* **31**: 587–589.

Cox CJ. 2018. Land plant molecular phylogenetics: a review with comments on evaluating incongruence among phylogenies. *Critical Reviews in Plant Sciences* **37**: 113–127.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Research* **14**: 1188–1190.

Diaz MF, Bentolila S, Hayes ML, Hanson MR, Mulligan RM. 2017. A protein with an unusually short PPR domain, MEF8, affects editing at over 60 *Arabidopsis* mitochondrial C targets of RNA editing. *The Plant Journal* **92**: 638–649.

Dombrovska E, Qiu Y-L. 2004. Distribution of introns in the mitochondrial gene *nad1* in land plants: phylogenetic and molecular evolutionary implications. *Molecular Phylogenetics and Evolution* **32**: 246–263.

Dong S, Wu H, Zhang S, Zhang L, Liu Y, Xue JY, Chen Z, Goffinet B. 2018. Complete mitochondrial genome sequence of *Anthoceros angustus*: conservative evolution of the mitogenomes in hornworts. *Bryologist* **121**: 14–22.

Edera AA, Gandini CL, Sanchez-Puerta MV. 2018. Towards a comprehensive picture of C-to-U RNA editing sites in angiosperm mitochondria. *Plant Molecular Biology* **97**: 1–17.

Freyer R, Kiefer-Meyer M-C, Kössel H. 1997. Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Sciences, USA* **94**: 6285–6290.

Grewe F, Herres S, Viehöver P, Polsakiewicz M, Weisshaar B, Knoop V. 2011. A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte *Isoetes engelmannii*. *Nucleic Acids Research* **39**: 2890–2902.

Groth-Malonek M, Pruchner D, Grewe F, Knoop V. 2005. Ancestors of *trans*-splicing mitochondrial introns support serial sister group relationships of hornworts and mosses with vascular plants. *Molecular Biology and Evolution* **22**: 117–125.

Guillaumot D, Lopez-Obando M, Baudry K, Avon A, Rigaill G, Falcon de Longevialle A, Broche B, Takenaka M, Berthomé R, De Jaeger G *et al.* 2017. Two interacting PPR proteins are major *Arabidopsis* editing factors in plastid and mitochondria. *Proceedings of the National Academy of Sciences, USA* **114**: 8877–8882.

Guo W, Grewe F, Mower JP. 2015. Variable frequency of plastid RNA editing among ferns and repeated loss of uridine-to-cytidine editing from vascular plants. *PLoS ONE* **10**: e0117075.

Guo W, Mower JP. 2013. Evolution of plant mitochondrial intron-encoded maturases: frequent lineage-specific loss and recurrent intracellular transfer to the nucleus. *Journal of Molecular Evolution* **77**: 43–54.

Haag S, Schindler M, Berndt L, Brennicke A, Takenaka M, Weber G. 2017. Crystal structures of the *Arabidopsis thaliana* organellar RNA editing factors MORF1 and MORF9. *Nucleic Acids Research* **45**: 4915–4928.

Hayes ML, Dang KN, Diaz MF, Mulligan RM. 2015. A conserved glutamate residue in the C-terminal deaminase domain of pentatricopeptide repeat proteins is required for RNA editing activity. *Journal of Biological Chemistry* **290**: 10136–10142.

Hayes ML, Giang K, Berhane B, Mulligan RM. 2013. Identification of two pentatricopeptide repeat genes required for RNA editing and zinc binding by C-terminal cytidine deaminase-like domains. *Journal of Biological Chemistry* **288**: 36519–36529.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biology and Evolution* **3**: 344–358.

Hein A, Brenner S, Knoop V. 2019. Multifarious evolutionary pathways of a nuclear RNA editing factor: disjunctions in co-evolution of DOT4 and its chloroplast target rpoC1eU488SL. *Genome Biology and Evolution* **11**: 798–813.

Hetherington AJ, Dolan L. 2018. Bilaterally symmetric axes with rhizoids composed the rooting structure of the common ancestor of vascular plants. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**: e20170042.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**: 518–522.

Ichinose M, Sugita M. 2018. The DYW domains of pentatricopeptide repeat RNA editing factors contribute to discriminate target and non-target editing sites. *Plant and Cell Physiology* **59**: 1652–1659.

Ichinose M, Sugita C, Yagi Y, Nakamura T, Sugita M. 2013. Two DYW subclass PPR proteins are involved in RNA editing of *ccmFc* and *atp9* transcripts in the moss *Physcomitrella patens*: first complete set of PPR editing factors in plant mitochondria. *Plant and Cell Physiology* **54**: 1907–1916.

Ichinose M, Uchida M, Sugita M. 2014. Identification of a pentatricopeptide repeat RNA editing factor in *Physcomitrella patens* chloroplasts. *FEBS Letters* **588**: 4060–4064.

Iyer LM, Zhang D, Rogozin IB, Aravind L. 2011. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Research* **39**: 9473–9497.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Knie N, Grewe F, Fischer S, Knoop V. 2016. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns – a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. *BMC Evolutionary Biology* **16**: e134.

Knoop V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Current Genetics* **46**: 123–139.

Kobayashi T, Yagi Y, Nakamura T. 2019. Comprehensive prediction of target RNA editing sites for PLS-class PPR proteins in *Arabidopsis thaliana*. *Plant and Cell Physiology* **60**: 862–874.

Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K. 2003a. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. *Nucleic Acids Research* **31**: 716–721.

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K. 2003b. RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Research* **31**: 2417–2423.

Kuraku S, Zmasek CM, Nishimura O, Katoh K. 2013. aLeaves facilitates on-demand exploration of metazoan gene family trees on Mafft sequence alignment server with enhanced interactivity. *Nucleic Acids Research* **41**: W22–W28.

Lenz H, Hein A, Knoop V. 2018. Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in Prepact 3.0. *BMC Bioinformatics* **19**: e255.

Lenz H, Rüdinger M, Volkmar U, Fischer S, Herres S, Grewe F, Knoop V. 2010. Introducing the plant RNA editing prediction and analysis computer tool Prepact and an update on RNA editing site nomenclature. *Current Genetics* **56**: 189–201.

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. Megahit v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**: 3–11.

Li L, Wang B, Liu Y, Qiu YL. 2009. The complete mitochondrial genome sequence of the hornwort *Megaceros aenigmaticus* shows a mixed mode of conservative yet dynamic evolution in early land plant mitochondrial genomes. *Journal of Molecular Evolution* **68**: 665–678.

Lohse M, Drechsel O, Kahlau S, Bock R. 2013. OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* **41**: W575–W581.

Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B *et al.* 2004. Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**: 2089–2103.

Lutzoni F, Nowak MD, Alfaro ME, Reeb V, Miadlikowska J, Krug M, Arnold AE, Lewis LA, Swofford DL, Hibbett D *et al.* 2018. Contemporaneous radiations of fungi and plants linked to symbiosis. *Nature Communications* **9**: e5451.

Malek O, Lättig K, Hiesel R, Brennicke A, Knoop V. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. *EMBO Journal* **15**: 1403–1411.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. 2018. The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences, USA* **115**: E2274–E2283.

Okuda K, Myouga F, Motohashi R, Shinozaki K, Shikanai T. 2007. Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. *Proceedings of the National Academy of Sciences, USA* **104**: 8178–8183.

Oldenkott B, Yamaguchi K, Tsuji-Tsukinoki S, Knie N, Knoop V. 2014. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. *RNA* **20**: 1499–1506.

Oldenkott B, Yang Y, Lesch E, Knoop V, Schallenberg-Rüdinger M. 2019. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in *Escherichia coli*. *Communications Biology* **2**: e85.

Peña MJ, Darvill AG, Eberhard S, York WS, O'Neill MA. 2008. Moss and liverwort xyloglucans contain galacturonic acid and are structurally distinct from the xyloglucans synthesized by hornworts and vascular plants. *Glycobiology* **18**: 891–904.

Piechotta M, Wyler E, Ohler U, Landthaler M, Dieterich C. 2017. Jacusa: site-specific identification of RNA editing events from replicate sequencing data. *BMC Bioinformatics* **18**: e7.

Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH, Schneider H, Pisani D *et al.* 2018. The interrelationships of land plants and the nature of the ancestral embryophyte. *Current Biology* **28**: 733–745.

Qiu YL, Cho Y, Cox JC, Palmer JD. 1998. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* **394**: 671–674.

Qiu Y-L, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L, Rest J *et al.* 2006. The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National Academy of Sciences, USA* **103**: 15511–15516.

Rensing SA. 2018a. Great moments in evolution: the conquest of land by plants. *Current Opinion in Plant Biology* **42**: 49–54.

Rensing SA. 2018b. Plant evolution: phylogenetic relationships between the earliest land plants. *Current Biology* **28**: R210–R213.

Rüdinger M, Funk HT, Rensing SA, Maier UG, Knoop V. 2009. RNA editing: only eleven sites are present in the *Physcomitrella patens* mitochondrial transcriptome and a universal nomenclature proposal. *Molecular Genetics and Genomics* **281**: 473–481.

Rüdinger M, Volkmar U, Lenz H, Groth-Malonek M, Knoop V. 2012. Nuclear DYW-type PPR gene families diversify with increasing RNA editing frequencies in liverwort and moss mitochondria. *Journal of Molecular Evolution* **74**: 37–51.

Salone V, Rüdinger M, Polsakiewicz M, Hoffmann B, Groth-Malonek M, Szurek B, Small I, Knoop V, Lurin C. 2007. A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Letters* **581**: 4132–4138.

Sandoval R, Boyd RD, Kiszter AN, Mirzakhanyan Y, Santibańez P, Gershon PD, Hayes ML. 2019. Stable native RIP9 complexes associate with C-to-U RNA editing activity, PPRs, RIPs, OZ1, ORRM1, and ISE2. *The Plant Journal* **99**: 1116–1126.

Schallenberg-Rüdinger M, Kindgren P, Zehrmann A, Small I, Knoop V. 2013. A DYW-protein knockout in *Physcomitrella* affects two closely spaced mitochondrial editing sites and causes a severe developmental phenotype. *The Plant Journal* **76**: 420–432.

Schallenberg-Rüdinger M, Knoop V. 2016. Coevolution of organelle RNA editing and nuclear specificity factors in early land plants. In: Rensing SA, ed. *Genomes and evolution of charophytes, bryophytes and ferns. Advances in Botanical Research, Vol. 78*. Amsterdam, the Netherlands: Elsevier Academic Press, 37–93.

Schultink A, Liu L, Zhu L, Pauly M. 2014. Structural diversity and function of xyloglucan sidechain substituents. *Plants* **3**: 526–542.

Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR. 2010. Extensive loss of RNA editing sites in rapidly evolving *Silene* mitochondrial genomes: selection vs. retroprocessing as the driving force. *Genetics* **185**: 1369–1380.

Smith DR. 2019. Unparalleled variation in RNA editing among *Selaginella* plastomes. *Plant Physiology*. doi: 10.1104/pp.19.00904.

de Sousa F, Foster PG, Donoghue PCJ, Schneider H, Cox CJ. 2018. Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist* **222**: 565–575.

Steinhauser S, Beckert S, Capesius I, Malek O, Knoop V. 1999. Plant mitochondrial RNA editing: extreme in hornworts and dividing the liverworts? *Journal of Molecular Evolution* **48**: 303–312.

Sugita M, Ichinose M, Ide M, Sugita C. 2013. Architecture of the PPR gene family in the moss *Physcomitrella patens*. *RNA Biology* **10**: 1439–1445.

Sun T, Bentolila S, Hanson MR. 2016. The unexpected diversity of plant organelle RNA editosomes. *Trends in Plant Science* **21**: 926–973.

Sun T, Germain A, Giloteaux L, Hammani K, Barkan A, Hanson MR, Bentolila S. 2013. An RNA recognition motif-containing protein is required for plastid RNA editing in Arabidopsis and maize. *Proceedings of the National Academy of Sciences, USA* **110**: E1169–E1178.

Sun T, Shi X, Friso G, Van Wijk K, Bentolila S, Hanson MR. 2015. A zinc finger motif-containing protein is essential for chloroplast RNA editing. *PLoS Genetics* **11**: e1005028.

Szövényi P, Frangedakis E, Ricca M, Quandt D, Wicke S, Langdale JA. 2015. Establishment of *Anthoceros agrestis* as a model species for studying the biology of hornworts. *BMC Plant Biology* **15**: e98.

Takenaka M, Zehrmann A, Brennicke A, Graichen K. 2013. Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. *PLoS ONE* **8**: e65343.

Takenaka M, Zehrmann A, Verbitskiy D, Kugelmann M, Härtel B, Brennicke A. 2012. Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. *Proceedings of the National Academy of Sciences, USA* **109**: 5104–5109.

Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. 2016. W-IQ-Tree: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* **44**: W232–W235.

Vangerow S, Teerkorn T, Knoop V. 1999. Phylogenetic information in the mitochondrial *nad5* gene of pteridophytes: RNA editing and intron sequences. *Plant Biology* **1**: 235–243.

Villarreal Aguilar JC, Turmel M, Bourgouin-Couture M, Laroche J, Salazar Allen N, Li F-W, Cheng S, Renzaglia K, Lemieux C. 2018. Genome-wide organellar analyses from the hornwort *Leiosporoceros dussii* show low frequency of RNA editing. *PLoS ONE* **13**: e0200491.

Villarreal JC, Forrest LL, Wickett N, Goffinet B. 2013. The plastid genome of the hornwort *Nothoceros aenigmaticus* (Dendrocerotaceae): phylogenetic signal in inverted repeat expansion, pseudogenization, and intron gain. *American Journal of Botany* **100**: 467–477.

Wagoner JA, Sun T, Lin L, Hanson MR. 2015. Cytidine deaminase motifs within the DYW domain of two pentatricopeptide repeat-containing proteins are required for site-specific chloroplast RNA editing. *Journal of Biological Chemistry* **290**: 2957–2968.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA *et al.* 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* **111**: E4859–E4868.

Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. Gmap and Gsnap for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology* **1418**: 283–334.

Xue JY, Liu Y, Li L, Wang B, Qiu YL. 2010. The complete mitochondrial genome sequence of the hornwort *Phaeoceros laevis*: retention of many ancient pseudogenes and conservative evolution of mitochondrial genomes in hornworts. *Current Genetics* **56**: 53–61.

Yagi Y, Hayashi S, Kobayashi K, Hirayama T, Nakamura T. 2013. Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS ONE* **8**: e57286.

Yan J, Yao Y, Hong S, Yang Y, Shen C, Zhang Q, Zhang D, Zou T, Yin P. 2019. Delineation of pentatricopeptide repeat codes for target RNA prediction. *Nucleic Acids Research* **47**: 3728–3738.

Yoshinaga K, Iinuma H, Masuzawa T, Uedal K. 1996. Extensive RNA editing of U to C in addition to C to U substitution in the *rbcL* transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. *Nucleic Acids Research* **24**: 1008–1014.

Zehrmann A, Härtel B, Glass F, Bayer-Császár E, Obata T, Meyer E, Brennicke A, Takenaka M. 2015. Selective homo and heteromer interactions between the Multiple Organellar RNA Editing Factor (MORF) proteins in *Arabidopsis thaliana*. *Journal of Biological Chemistry* **290**: 6445–6456.

Zumkeller SM, Knoop V, Knie N. 2016. Convergent evolution of fern-specific mitochondrial group II intron atp1i361g2 and its ancient source paralogue rps3i249g2 and independent losses of intron and RNA editing among Pteridaceae. *Genome Biology and Evolution* **8**: 2505–2519.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 The tRNA-Asp(GUC) as an example for mitochondrial RNA editing in a structural RNA.

Fig. S2 Three 'DYW-only'-type PPR proteins of *Anthoceros agrestis*.

Table S1 *Anthoceros agrestis* chloroplast RNA editing sites.

Table S2 *Anthoceros agrestis* mitochondrial RNA editing sites.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.